Preliminaries to an Account of Multi-Party Conversational Turn-Taking as an Antiferromagnetic Spin Glass

Kornel Laskowski, Mattias Heldner & Jens Edlund[*]
KTH Speech Music and Hearing, Stockholm, Sweden
`kornel@cs.cmu.edu`

Abstract

We present empirical justification of why logistic regression may acceptably approximate, using the number of currently vocalizing interlocutors, the probabilities returned by a time-invariant, conditionally independent model of turn-taking. The resulting parametric model with 3 degrees of freedom is shown to be identical to an infinite-range Ising antiferromagnet, with slow connections, in an external field; it is suitable for undifferentiated-participant scenarios. In differentiated-participant scenarios, untying parameters results in an infinite-range spin glass whose degrees of freedom scale as the square of the number of participants; it offers an efficient representation of participant-pair synchrony. We discuss the implications of model parametrization and of the thermodynamic and feed-forward perceptron formalisms for easily quantifying aspects of conversational dynamics.

1 Introduction

The organization of conversational turn-taking, particularly in multi-participant scenarios, eludes efforts at straighforward, tractable modeling, despite claims that conversations "exhibit their orderliness, have their orderliness appreciated and used, and have that appreciation displayed and treated as the basis for subsequent action" [15].

The stochastic modeling of multi-participant *chronograms* [3] of speech activity as vector-valued Markov processes is one of several alternatives [18]. A chronogram $\mathbf{Q} \equiv \{\mathbf{q}_1, \ldots, \mathbf{q}_T\}$ is a representation of conversation which elides lexical context, non-durational prosody, and other non-vocal behavior. Each K-length vector $\mathbf{q}_t$, $1 \leq t \leq T$, concatenates the binary vocal activity state $\in \mathbb{S} \equiv \{\square, \blacksquare\}$ at instant t of all participants. It is therefore drawn from the set $\mathbb{S}^K \equiv \{\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_{N-1}\}$, of $N = 2^K$ alternatives. Chronograms describe only the relative *timing* of the deployment of vocal activity in conversation; that modeling them stochastically does not require a definition of what a *turn* might actually be is in our view its main strength.

The goal of a general-purpose, time-independent turn-taking model under these assumptions is to provide the prior likelihood of $\mathbf{Q}$. With first-order history truncation,

$$P(\mathbf{Q}) \quad \dot{=} \quad P(\mathbf{q}_0) \prod_{t=1}^{T} P(\mathbf{q}_t \mid \mathbf{q}_{t-1}) . \tag{1}$$

A constraint of *conditional independence* among participants' states can be implied by further factoring each bigram factor in Equation 1 into

$$P(\mathbf{q}_t \mid \mathbf{q}_{t-1}) \quad \dot{=} \quad \prod_{k=1}^{K} P(\mathbf{q}_t[k] \mid \mathbf{q}_{t-1}[k], \mathbf{C}_k \mathbf{q}_{t-1}) , \tag{2}$$

where $\mathbf{C}_k$ is the $K \times K$ identity matrix with the kth column removed. This form, known as the "single-port" model in [1] and the "separate source" model in [6], has recently been implemented in a spoken dialogue system [14].
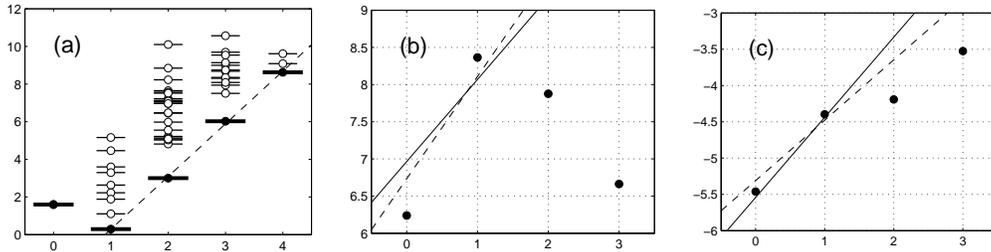
Figure 1: A meeting from the ICSI Meeting Corpus [7], `Bmr025`, with $K = 8$ participants and $T \approx$ 20K 100-ms frames. (a) Negative log-probabilities (along y-axis) of observed states $\mathbf{S}_i$, as a function of the degree $\|\mathbf{S}_i\|$ of overlap in that state (along x-axis); dashed line indicates the unweighted least-squares fit to $-\log P(\|\mathbf{S}_i\|)$ as a function of $\|\mathbf{S}_i\|$. (b) & (c) Negative conditional logit-probabilities (along y-axes) of initiating and continuing speech, respectively, at instant t given the degree of overlap observed among interlocutors (along x-axes) at instant $t-1$. Dashed line indicates weighted least-squares fit to $-\mathrm{logit}P(\blacksquare|\square, \|\mathbf{C}_k\mathbf{S}_i\|)$ and $-\mathrm{logit}P(\blacksquare|\blacksquare, \|\mathbf{C}_k\mathbf{S}_i\|)$, respectively. Solid line indicates a fit with the constraint that the slopes in (b) and (c) be identical. Weighting achieved using $P(\|\mathbf{C}_k\mathbf{S}_i\|)$ in both cases. These trends are observed for all ICSI meetings.

In *multi-party* conversation, prediction of vocal activity using direct estimates of the probabilities in Equations 1 and 2 has been shown to be infelicitous for several reasons [9], chief among them being the size of the state space $\mathbb{S}^K$. To address this problem, we have recently proposed [10] a conditionally independent "single-port" variant of the EDO model of [9], which ties conditioning contexts $\mathbf{C}_k\mathbf{q}_{t-1}$ by their *degree of overlap*, or the number of participants vocalizing simultaneously in them, $\|\mathbf{C}_k\mathbf{q}_{t-1}\|$. The SPEDO model yields

$$P(\mathbf{q}_t[k] \mid \mathbf{q}_{t-1}[k], \mathbf{C}_k\mathbf{q}_{t-1}) \doteq P(\mathbf{q}_t[k] \mid \mathbf{q}_{t-1}[k], \|\mathbf{C}_k\mathbf{q}_{t-1}\|) \qquad (3)$$

and is the point of departure for the current argument. The number of independent parameters in this model is $2K$. Consequently, longer n-grams have $(2K)^{(n-1)}$ independent parameters; our experience [10] is that SPEDO models begin to underperform when their number of parameters, estimated from over 60 hours of meetings, exceeds approximately 200. This circumscribes the length of history which can be used to condition estimates. A common aspect of stochastic turn-taking models is that they are *non-parametric*; each probability which the models return must be separately modeled and stored (less one per state, since they sum to unity). The current work describes the justification for a *parametric* model, in which the required probabilities are generated from an "internal" model, whose own parameters are fewer in number and are not those probabilities themselves.

2 Ancillary Observations on Overlap

In [17], it was reported that 31.4–54.4% of talkspurts in a large meeting corpus exhibit some overlap; similar numbers have been reported for dialogue [4]. On a still finer scale, [2] showed using a multi-site multi-party corpus that 11.6% of speech by time exhibits overlap, and that of all overlap time, 92.2% consists of simultaneous vocalization by only 2 participants. We extend those analyses here, by partitioning all overlap by degree. This exposes the relationship between the degree $\|\mathbf{S}_i\|$ of overlap in state $\mathbf{S}_i$ and the negative log-likelihood of that state's occurrence in a conversation $\mathbf{Q}$ of duration T, $-\log P(\mathbf{S}_i) = -\log \frac{1}{T}\sum_{t=1}^{T} \delta_K(\mathbf{q}_t, \mathbf{S}_i)$; where δ_K is defined as the Kronecker delta extended to K-length vector arguments. Figure 1 depicts the points $(\|\mathbf{S}_i\|, -\log P(\mathbf{S}_i))$ for all states observed in a single meeting as short horizontal lines. The figure also shows the points $(\|\mathbf{S}_i\|, -\log P(\|\mathbf{S}_i\|))$, where $-\log P(\|\mathbf{S}_i\|) = -\log \frac{1}{T}\sum_{t=1}^{T} \delta(\|\mathbf{q}_t\|, \|\mathbf{S}_i\|)$, with longer horizontal lines. These latter points indicate the probability of being in *any* state with the same degree of overlap. The probability of any non-zero degree of overlap is surprisingly easy to predict.

3 Logistic Regression

Figure 1(a) depicts the *unconditional* (unigram) probability of contrastive degrees of overlap, whereas Equation 3 requires the degree $\|\mathbf{C}_k\mathbf{q}_{t-1}\|$ of overlap only among *interlocutors* of par-

ticipant k to be modeled, in the *conditioning context* of a Bernoulli variable $\mathbf{q}_t[k]$. Because $P(\mathbf{q}_t[k] = \blacksquare|\ldots) = 1 - P(\mathbf{q}_t[k] = \square|\ldots)$, we treat only *initiating vocalization* and *continuing vocalization*, $P(\blacksquare|\square, \|\mathbf{C}_k\mathbf{q}_{t-1}\|)$ and $P(\blacksquare|\blacksquare, \|\mathbf{C}_k\mathbf{q}_{t-1}\|)$, respectively.

Panels (b) and (c) of Figure 1 depict these probabilities, inferred from the same meeting used to produce panel (a), as a function of the degree of interlocutor overlap. They are shown on the negative *logit* scale, rather than the negative log scale, since a linear fit on the latter may lead to probability estimates outside the unit interval; the scales are visually identical for small probability values. As can be seen in panel (b), the probability of initiating vocalization is larger when two interlocutors are observed to be vocalizing than when only one is; the same is true for three interlocutors relative to two. Evidently, the floor is more "up for grabs" when more than one person is already seen to be vocalizing. The least squares fit, weighted by the likelihood of interlocutor overlap and shown as a dashed line, poorly predicts behavior when $\|\mathbf{C}_k\mathbf{q}_{t-1}\| > 1$ (this accounts for $< 5\%$ of frames).

A similarly produced fit for continuing vocalization is much better, as evidenced in panel (c). The monotonic fall in probability of continuing to vocalize in overlap is described in conversation analysis as "Talk by MORE than two at a time seems to be reduced to two (or to one) even more efficiently than talk by two is reduced to one" [16]. A weighted least squares fit, constrained such that the slope of the fit in panels (b) and (c) is the same, is shown as a solid line in both panels; it does not deviate from the dashed lines egregiously around the points with the largest $P(\|\mathbf{C}_k\mathbf{q}_{t-1}\|)$ weights.

We now assign to the common slope the symbol w_-, and note that the solid lines in both panels are described by $-\mathrm{logit}(P(\ldots)) \equiv -h_t = b + w_+\mathbf{q}_{t-1}[k] + w_-\sum_{j \neq k}\mathbf{q}_{t-1}[j]$, with $\{\square, \blacksquare\}$ coded as $\{0, 1\}$. Alternately, in matrix notation, $\mathbf{h}_t = -(b + \mathbf{W}\mathbf{q}_{t-1})$. $\mathbf{W}$ is a $K \times K$ matrix with w_+ along the diagonal and w_- elsewhere. The probability that participant k is in state $\blacksquare$ is then simply $P(\mathbf{q}_t[k] = \blacksquare|\mathbf{q}_{t-1}) = 1/(1 + e^{-\beta\mathbf{h}_t[k]})$, the inverse of the logit function. β^{-1} is the pseudo-temperature, and is clamped to unity here. This formulation makes $\mathbf{W}$ the connectivity matrix of an *Ising model* [13]. The absence of off-diagonal zeros makes it an *infinite range* model; each participant influences every other. That all off-diagonal entries are positive suggests that the model is *anti*ferromagnetic; its minima are attained when only one person vocalizes at a time. The non-zero diagonal entries denote what are known as "slow" dynamic connections [5].

4   Some Consequences of This Account

The above formulation yields a 3-parameter model of multi-party turn-taking, which assumes that all participants are governed by identical dynamics. This is parsimonious relative to other models; the SPEDO model, which the Ising model approximates, has $2K$ free parameters [10]. Presently, we are investigating methods of learning $\{b, w_-, w_+\}$ *across* conversations, to be applied to *unseen* conversations; b appears to depend on K, and we expect $w_+ + b < 0$ for sub-talkspurt frame steps.

Modeling the *future* within conversations, over the course of which neither the number nor index assignment of participants changes, can be improved by relaxing the undifferentiated-participant tying in $\mathbf{W}$. When all $K \times K$ parameters are free, $\mathbf{W}$ becomes the connectivity matrix of a *spin glass*. The value in row k and column j indicates the extent to which participant j inhibits participant k from speaking. Inference of $\mathbf{W}$ can be achieved via the *reweighted least squares* algorithm, or, as we prefer in anticipation of future extensions, *backpropagation* in the *cross-entropy error*. This is because $\mathbf{W}$ and $\mathbf{b}$ also implement the weights and biases of a single-layer perceptron with sigmoid activations and no hidden units. We have estimated $\mathbf{W}$ this way to classify conversation type [11], as well as participant role and seniority [12]. Second pass estimation of a time-varying pseudo-temperature, with $\mathbf{W}$ and $\mathbf{b}$ fixed, has been used to locate conversational hot spots [8].

We are planning to port the Ising and spin glass models described above to an online system. There are three aspects of the current argument which deserve specific mention. First, the parametrization of the discrete symbol SPEDO model facilitates the inclusion of any feature domain in the conditioning context. We are particularly interested in augmenting the model with intonation, loudness, and syllable rate features. Provided that features can be computed without prior segmentation, the parametric models can in principle be used to predict future vocal activity without having to segment past vocal activity. Second, the thermodynamic formalism offers many novel tools for studying conversations as general K-body dynamic systems. The design of a temperature measure to quantify "heated" departure from turn-taking norms, and to characterize instances and varieties of conversa-

tion, is one example. Finally, the neural network formalism enables flexible combination and coding of transition probabilities, as well as myriad ways to improve regression via hidden units (but with a larger parameter space). We anticipate that it will also be relatively easy to extend the models with softmax outputs to predict different types of vocal activity, including laughter and backchanneling.

5 Conclusions

This paper has presented a justification of employing a parsimonious and well-understood formalism, with roots in both statistical physics and computational neuroscience, to the problem of modeling the taking of turns in multi-party conversation. The justification consists of the discovery of a near-linear relationship between a multi-participant state's negative log-likelihood of occurrence and its degree of overlap, and similar but weaker trends governing conditional probabilities with interlocutor overlap in the conditioning context. The exploitation of this relationship yields naturally a family of extensions of the Ising model. We have mentioned our past successes with this model, our current work on parameter regression across conversations, and our immediate plans to implement the model in spoken dialogue system settings and experiments.

References

[1] P. T. BRADY, *A model for generating on-off speech patterns in two-way converation*, Bell Systems Technical Journal, 48 (1969), pp. 2445–2472.

[2] O. 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, pp. 212–224.

[3] E. CHAPPLE, *The interaction chronograph: Its evolution and present application*, Personnel, (1949), pp. 295–307.

[4] M. HELDNER AND J. EDLUND, *Pauses, gaps and overlaps in conversations*, Journal of Phonetics, (2010). doi:10.1016/j.wocn.2010.08.002.

[5] J. HERTZ, A. KROGH, AND R. G. PALMER, *Introduction to the Theory of Neural Computation*, Perseus Books Publishing, Reading MA, USA, 1991.

[6] J. JAFFE AND S. FELDSTEIN, *Rhythms of Dialogue*, Academic Press, New York NY, USA, 1970.

[7] A. JANIN, D. BARON, J. EDWARDS, ET AL., *The ICSI Meeting Corpus*, in Proc. ICASSP, Hong Kong, China, 2003, pp. 364–367.

[8] K. LASKOWSKI, *Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings*, in Proc. SLT, Goa, India, 2008, pp. 81–84.

[9] ——, *Modeling norms of turn-taking in multi-party conversation*, in Proc. ACL, Uppsala, Sweden, 2010, pp. 999–1008.

[10] K. LASKOWSKI, M. HELDNER, AND J. EDLUND, *A single-port non-parametric model of turn-taking in multi-party conversation*, in Proc. ICASSP, Praha, Czech Republic, 2011. submitted.

[11] K. LASKOWSKI, M. OSTENDORF, AND T. SCHULTZ, *Modeling vocal interaction for text-independent classification of conversation type*, in Proc. SIGdial, Antwerpen, Belgium, 2007, pp. 194–201.

[12] ——, *Modeling vocal interaction for text-independent participant characterization in multi-party meetings*, in Proc. SIGDIAL, Columbus OH, USA, 2008, pp. 148–155.

[13] D. J. C. MACKAY, *Information Theory, Inference, and Leaning Algorithms*, Cambridge University Press, Cambridge, UK, 2003.

[14] A. RAUX AND M. ESKENAZI, *Finite state turn taking model for spoken dialog systems*, in Proc. HLT-NAACL, Boulder CO, USA, 2009, pp. 629–637.

[15] H. SACKS, E. SCHEGLOFF, AND G. JEFFERSON, *A simplest semantics for the organization of turn-taking for conversation*, Language, 50 (1974), pp. 696–735.

[16] E. A. SCHEGLOFF, *Overlapping talk and the organization of turn-taking for conversation*, Language in Society, 29 (2000), pp. 1–63.

[17] 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, pp. 1359–1362.

[18] T. P. WILSON, J. M. WIEMANN, AND D. H. ZIMMERMAN, *Models of turn taking in conversational interaction*, Journal of Language and Social Psychology, 3 (1984), pp. 159–183.