A SINGLE-PORT NON-PARAMETRIC MODEL OF TURN-TAKING IN MULTI-PARTY CONVERSATION

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

[1] KTH Speech, Music and Hearing, Stockholm, Sweden
[2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA

ABSTRACT

The taking of turns to speak is an intrinsic property of conversation. It is expected that models of taking turns, providing a prior distribution over conversational form, can reduce the perplexity of what is attended to and processed by spoken dialogue systems. We propose a single-port model of multi-party turn-taking which allows conversants to behave independently but to condition their behavior on the past of the entire group. The model performs at least as well as an existing multi-port model on perplexity over subsequent speech activity. We quantify the effect of longer histories and more distant future horizons, and argue that the framework has the potential to inform the design and behavior of spoken dialogue systems.

Index Terms— turn-taking, speech processing, ngram modeling, time series prediction, dialogue systems.

1. INTRODUCTION

The on-off pattern of speech activity in human-human conversation [1] — its precise distribution in time and across participants — is said to be grossly accountable for by a simple systematics of *turn-taking* [2]. The computational modeling of this phenomenon has recently seen a resurgence of interest, for two apparent reasons. First, services are increasingly being offered by synthetic agents with spoken dialogue interfaces. Attempts to make interaction with such agents more human-like has put prediction with regard to turn-taking in the spotlight [3]. Second, the need to automatically extract information from human-human conversation, particularly multi-party meetings, calls for a detection framework which explicitly licenses the production of speech in overlap [4].

If the state of a conversation is defined as the concatenation of the speech/non-speech states of its participants, then both the prediction and the detection tasks can be addressed by a single model, one yielding a conditional probability measure over alternative conversational futures. (A product of such measures over consecutive instants provides a likelihood density estimate of the conversation in the circumscribed interval.) Additionally assuming the process which generates conversations to be Markovian renders parameter estimation tractable. The resulting n-gram form, popularized by language modeling in speech recognition, has been applied extensively to two-party conversation [5, 1, 6, 7].

In more-than-two-party conversation, Markov modeling of the multi-participant speech activity state was explored in [8], but only under the assumption of conditional dependence among participants. While this assumption poses no difficulties for detection, the same is not true for prediction. Dialogue systems may wish to evaluate the potential consequences of their planned actions, without knowledge of what others are planning. Doing so with a conditionally dependent form requires that others' speech activity states be marginalized out; when the number of others is large, the process can be operationally cumbersome and needlessly time-consuming.

The aim of this paper is three-fold. First, a model is developed which yields the probability of specific combinations of participants speaking at instant t (as in [8]), but under the assumption that they behave independently given their joint, multi-participant speech/non-speech state (■/□) at instant $t-1$. This is known as the ""independent decision" hypothesis" in [5], the "single-port" (versus "multi-port") model in [1], and the "separate source" (versus — somewhat confusingly — "single source") model in [6]. Second, the model is evaluated for its ability to limit the perplexity of speech activity observed in naturally occurring multi-party conversations, using the framework in [8]. Third, the paper explores n-gram truncations for $n > 1$. Collectively, the findings provide a convenient means of collaborative, model-based, fine-grained synthesis of speech/non-speech patterns for conversations of arbitrary duration and of arbitrary participant number.

2. DATA

Analysis and experiments are performed using the ICSI Meeting Corpus [9, 10]. 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, yielding a wide variety of interaction types. The total meeting time in the corpus is 67 hours.

All experiments are conducted in leave-one-out round-robin fashion, in which, for each meeting, predictive models are constructed using the remaining meetings.

3. BASELINE MODELS

This paper compares directly to the work in [8], where the thrust was to estimate the perplexity

$$\text{PPL} \quad = \quad \left(P\left(\,\mathbf{Q}\,|\,\mathbf{\Theta}\,\right)\right)^{-1/KT} \tag{1}$$

of the vocal interaction chronogram $\mathbf{Q}$ of a conversation, unseen and unanticipated during the training of the model $\mathbf{\Theta}$. The chronogram [11] is the frame-synchronous speech/non-speech segmentation of all the participants to a conversation,

$$\mathbf{Q} \quad \equiv \quad [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_T] \quad \in \quad \{\square \equiv 0, \blacksquare \equiv 1\}^{K \times T} \,, \tag{2}$$

where $K > 2$ is the known but arbitrary number of participants and T is the known but arbitrary number of 100-ms frames. The

likelihood in Equation 1 is given by factoring $\mathbf{Q}$,

$$P(\mathbf{Q}) = P_0 \prod_{t=1}^{T} P(\mathbf{q}_t \mid \mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_{t-1}) \quad (3)$$

$$\doteq P_0 \prod_{t=1}^{T} P(\mathbf{q}_t \mid \mathbf{q}_{t-1}) \ . \quad (4)$$

Equation 4 makes the standard 1st-order Markov assumption; P_0 is the unigram probability of an artificially pre-pended "all silent" state which does not affect our bigram model comparisons. Each factor in the remaining product, in [8], was given by the Extended Degree-of-Overlap (EDO) model,

$$P(\mathbf{q}_t \mid \mathbf{q}_{t-1}) = \alpha P(\|\mathbf{q}_t\|, \|\mathbf{q}_t \cdot \mathbf{q}_{t-1}\| \mid \|\mathbf{q}_{t-1}\|) \ , \quad (5)$$

where $\|\mathbf{q}\| \equiv \sum_{k=1}^{K} \mathbf{q}[k]$ yields the number of participants in the ■ state in $\mathbf{q}$, and $\mathbf{q} \cdot \mathbf{q}'$ is the element-wise logical AND. It yields a K-length vector whose kth entry is ■ if and only if the kth entries of both $\mathbf{q}$ and $\mathbf{q}'$ are ■. α is a normalization constant which first distributes probability mass among transitions with identical EDO transition types, and then normalizes the sum of transition probabilities out of each state to unity.

The EDO model allows for the imposition of a maximum degree K_{max} of modeled overlap, with higher, actually occurring degrees mapped onto K_{max} (cf. Equations 18–20 in [8]). While this mapping was applied during model training in [8], it was not applied during scoring. A consequence is that the unseen-conversation perplexities reported in [8] are lower than in actuality[1]. Both the original and the corrected perplexities are shown in Figure 1.

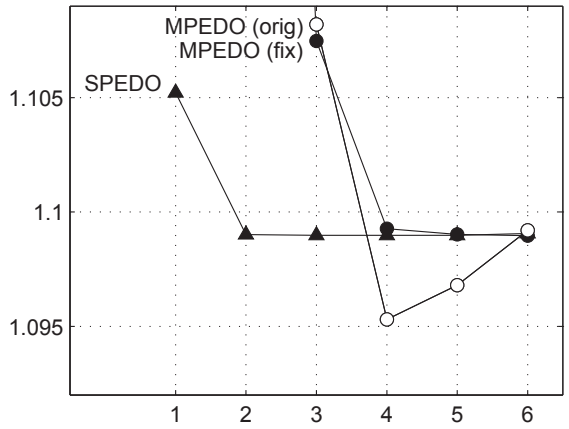

Fig. 1. Perplexity (along y-axis) as a function of $K_{max} \in \mathbb{Z}$ (along x-axis). The originally reported and then corrected multi-port EDO model perplexities ("MPEDO (orig)" and "MPEDO (fix)", respectively) are shown alongside the perplexities of the newly proposed single-port EDO model. Lines connecting points are provided for visualization purposes only.

[1]The reported [8] reduction of perplexity from a mutually independent (single-participant) model, achieving a PPL $= 1.1051$, to the best corrected EDO model (PPL $= 1.0990$ at $K_{max} = 6$), where the minimum of PPL $= 1.0921$ is given by an "oracle" conditionally dependent model, is 47% rather than 75%. Similarly, when only instants for which $\mathbf{q}_{t-1} \neq \mathbf{q}_t$ are considered, the reduction is $(1.8170 - 1.7380) \,/\, (1.8170 - 1.6616) = 51\%$ rather than 78%.

4. CONDITIONAL INDEPENDENCE

As discussed briefly in the introduction, the assumption of conditional dependence is not quite appropriate in the context of interacting conversants. It entails accepting that participants first negotiate and jointly agree on how each of them will behave *at the next instant*, and only then proceed by executing the agreed upon action(s). A more plausible account is that participants make their own judgments as to whether or not to vocalize (i.e. act independently, conditioned on the joint past), observe the joint outcome at the next instant, and *resolve* any conflicts that have occurred with respect to the floor. This is in fact the mechanism proposed to explain overlap dynamics in conversation analysis [12].

The mechanism can be easily formalized as

$$P(\mathbf{Q}) = P_0 \prod_{t=1}^{T} \prod_{k=1}^{K} P(\mathbf{q}_t[k] \mid \mathbf{q}_{t-1}[k], \mathbf{C}_k \mathbf{q}_{t-1}) \ , \quad (6)$$

where $\mathbf{C}_k$ is a matrix obtained by eliminating the kth column from the $K \times K$ identity matrix $\mathbf{I}$. We have chosen to split the multi-participant conditioning context $\mathbf{q}_{t-1}$ to clearly identify the effect of one's own past behavior and that of one's participants.

Estimating the probabilities in Equation 6 directly leads to models which are specific to the number of participants (through the size of the state vectors $\mathbf{q}$) and to the index assignment of participants in $\mathbf{q}$. They are therefore suitable only when applied to (potentially unseen portions of) the same conversations on which they are trained (cf. [8], where the direct compositional conditionally independent model, with $K \cdot 2^K$ free parameters, is identified as $\{\mathbf{\Theta}_k^{CI}\}$).

The main contribution of the current work is to propose an alternative conditionally independent form for Equation 6 which is not specific to participant k, or to the number of participants K, or to the index assignment of participants in $\mathbf{Q}$. We refer to the proposal as the single-port extended degree-of-overlap (SPEDO) model,

$$P(\mathbf{Q}) = P_0 \prod_{t=1}^{T} \prod_{k=1}^{K} P(\mathbf{q}_t[k] \mid \mathbf{q}_{t-1}[k], \|\mathbf{C}_k \mathbf{q}_{t-1}\|) \ . \quad (7)$$

Rather than conditioning on previous vocal activity states of specific interlocutors, given by the elements of the ordered vector $\mathbf{C}_k \mathbf{q}_{t-1}$, only the *degree of interlocutor overlap*, $\|\mathbf{C}_k \mathbf{q}_{t-1}\|$, is used. (An *unconditionally independent* model, popular in the acoustic detection of vocal activity, is obtained by eliminating $\|\mathbf{C}_k \mathbf{q}_{t-1}\|$ from the conditioning context altogether.)

5. EXPERIMENTS

5.1. Differentiating Among Degrees of Overlap

To more finely control the state space of the model, we introduce the ceiling K_{max}, with $K_{max} - 1$ indicating the maximum differentiable number of interlocutors vocalizing simultaneously. For example, if $K_{max} = 4$, then the model can only differentiate between conditioning contexts of zero, one, two, or *three-or-more* simultaneously vocalizing interlocutors; it cannot differentiate among degrees of interlocutor overlap of three or four (or higher).

Perplexities achieved by a SPEDO bigram model, with $K_{max} \in \{1, 2, 3, 4, 5, 6\}$, are shown in Figure 1. $K_{max} = 0$ corresponds to ignoring interlocutor vocal activity; $K_{max} = 1$ corresponds to being sensitive to either zero, or *any* non-zero number of interlocutors vocalizing; $K_{max} = 2$ corresponds to being sensitive to either zero,

exactly one, or any number greater than one of interlocutors vocalizing; etc. Also shown are the perplexities from the multi-port EDO baseline (MPEDO), in its original and corrected form, for which K_{max} refers to the total number of participants speaking, rather than the total number of interlocutors speaking (seen from the point of view of any single participant).

The diagram makes clear that the SPEDO model is generally better than the corrected MPEDO model, with the latter doing particularly poorly for $K_{max} < 4$ as already shown in [8]. The MPEDO model is only negligibly better for $K_{max} = 6$.

5.2. Conditioning on Longer Truncations of History

Whereas bigram models predict what happens at instant t given only what happened at instant $t-1$, it is methodologically straightforward to assess whether longer conditioning histories improve predictions, by employing longer n-grams,

$$P(\mathbf{Q}) \doteq P_0 \prod_{t=1}^{T} P\left(\mathbf{q}_t \mid \mathbf{q}_{t-(n-1)}, \ldots, \mathbf{q}_{t-1}\right) . \quad (8)$$

We propose the following factor expansion as a SPEDO counterpart:

$$\begin{aligned} & P\left(\mathbf{q}_t[k] \mid \mathbf{q}_{t-(n-1)}, \ldots, \mathbf{q}_{t-1}\right) \quad (9) \\ &= P\left(\mathbf{q}_t[k] \mid \mathbf{q}_{t-(n-1)}[k], \|\mathbf{C}_k \mathbf{q}_{t-(n-1)}\|, \right. \\ &\qquad \mathbf{q}_{t-(n-2)}[k], \|\mathbf{C}_k \mathbf{q}_{t-(n-2)}\|, \\ &\qquad \left. \ldots, \mathbf{q}_{t-1}[k], \|\mathbf{C}_k \mathbf{q}_{t-1}\|, \right) . \end{aligned}$$

That is, for each instant of history, we model both the state of the participant in question *and* that participant's number of vocalizing interlocutors. The number of free parameters for general n is $(2K)^{n-1}$. Figure 2 presents our experimental results; the perplexities shown are those achieved without back-off or smoothing.

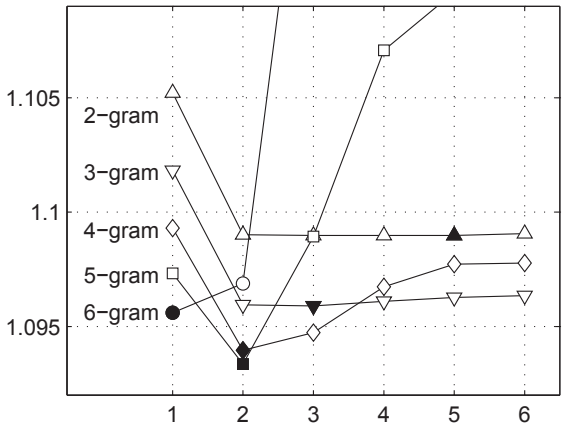

Fig. 2. Perplexity (along y-axis) as a function of $K_{max} \in \mathbb{Z}$ (along x-axis), for 1st- through 5th-order Markov truncated models. Lines connecting points are provided for visualization purposes only. Points corresponding to each n-gram model's K_{max} value which yields the lowest perplexity for that model's Markov order are shown in black.

It can be seen that the performance of the trigram, which conditions predictions on behavior at $t-1$ and $t-2$, is better than for the bigram for all K_{max}. Evidently, longer histories are beneficial. However, the lowest perplexities for the trigram are achieved when K_{max} is equal to 3, where only degrees of interlocutor overlap of zero, one, and two-or-more are differentiated. Subsequent increases in K_{max} are accompanied by *higher* perplexity, suggesting that the model begins overfitting (at $K_{max} = 3$, the number of free parameters is 36). The 4-gram begins overfitting earlier, at $K_{max} = 2$, thus differentiating only among zero or one-or-more simultaneously vocalizing interlocutors. Nevertheless, the additional context at $t-3$ leads to lower perplexity than does the sensitivity to one additional degree of interlocutor overlap. This trend continues for the 5-gram, which achieves the lowest observed perplexity, also at $K_{max} = 2$. The 6-gram, better than the 5-gram when interlocutors are ignored ($K_{max} = 1$), never outperforms the 5-gram model for $K_{max} > 1$.

5.3. Predicting Further into the Future

Equally interesting is whether the proposed modeling technique can be used to make predictions further into the future than merely at the *immediately next* instant. This is easily verified using the *skip n-gram* formalism, which provides for what happens at $t+m$, $m > 0$, rather than at t,

$$\begin{aligned} & P\left(\mathbf{q}_{t+m}[k] \mid \mathbf{q}_{t-1}[k], \mathbf{C}_k \mathbf{q}_{t-1}, \ldots\right) \quad (10) \\ & \qquad \doteq P\left(\mathbf{q}_{t+m}[k] \mid \mathbf{q}_{t-1}[k], \|\mathbf{C}_k \mathbf{q}_{t-1}\|, \ldots\right) \end{aligned}$$

with any desired duration of the conditioning history; m is the number of immediately future instants which are "skipped". It is important to note that, regardless of how far back this model is allowed to look (controlled by n), it predicts vocal activity at only one instant into the future, m instants ahead.

The experiments in this section, summarized in Figure 3, use at each K_{max} that n-gram model from Figure 2 which achieves the lowest perplexity for *that* value of K_{max} specifically (these points are those labeled "skip-0" in Figure 3). We then retrain that fixed-n model for $m \in \{1, 2, 3, 4\}$. Four 100-ms frames correspond to two to three average-duration syllables, or approximately one word.

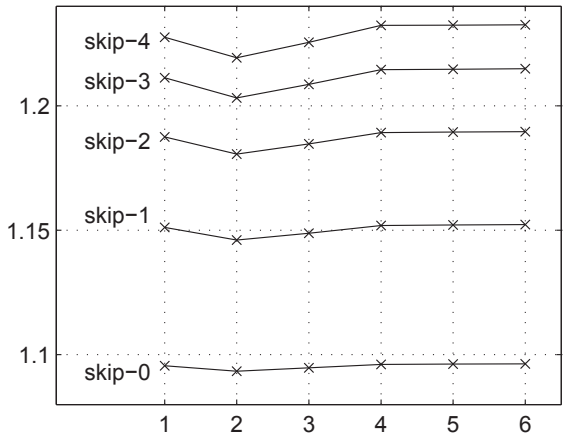

Fig. 3. Perplexity (along y-axis) as a function of $K_{max} \in \mathbb{Z}$ (along x-axis), for skip-m n-gram models; n is chosen by selecting that which was best-performing for each K_{max} in Figure 2. Lines connecting points are provided for visualization purposes only.

Since the frame-level perplexity measure was only recently introduced, the y-scale is difficult to interpret. It is therefore problematic to assess the impact of the results in Figure 3, except that

a perplexity of 2 would characterize a model always guessing fifty-fifty if the prior over $(\square, \blacksquare)$ was also fifty-fifty. Two points can be made, however. First, as expected, same-complexity models yield poorer and poorer predictions of an increasingly distant horizon of m. Second, even for $m = 1$, perplexities are much higher than for any of the models of Figure 2, where m was uniformly zero.

It is important to note that predicting what might happen at $m > 0$ is not only useful for modeling the *auxillary* capabilities of more "forward-looking" conversational partners, but is also quite likely a suitable paradigm for modeling the *primary* capacity of "slower-to-respond" (or simply "stubborn") partners. The latter may merely not observe (or may choose not to react to) what their interlocutors were doing as recently as at $t-1$ until yet another instant has passed by, effectively inserting a temporal response gap between what is observed and what can be predicted at the next instant.

6. DISCUSSION

The proposed framework provides an efficient setting for subjective modeling of a salient aspect — the on-off speech patterns — of each participant's conversational behavior. Whereas the discussion has treated a single, participant-independent model, the extension to training several dissimilar models using real data is straightforward. Such models could be used as a basis for simulating emergent group behaviors as a function of participant tendencies, or "personalities", in an organic bottom-up fashion. For example, simulation using models trained on the most talkative individuals in real conversations is expected to yield drastically different results from that based on only the least talkative people.

For spoken dialogue systems that have something to say, the models can predict at what future instant it is appropriate to start speaking. Arguably, even a system that does not have something to say, for example one that is awaiting information from a database search, can start producing floor holders at the point in time that the model suggests — particularly if the system's models of *its* interlocutors indicate that those interlocutors might otherwise begin speaking, thereby taking from it the initiative. Investigative methods employing such reasoning may lead to systems which are perceived as more polite, efficient and generally better at turn-taking [13]. There is scope for changing system personality, which is highly interesting for constructing artificial conversational partners.

Finally, the models are sufficiently parsimonious to be manually approachable, "by hand". A bigram model with $K_{max} = 3$ consists of only 6 Bernoulli probabilities, each of which can be individually tweaked in order to analyze its impact on unfolding conversational patterns. The probabilities can be interpolated with those obtained by additionally modeling other features, such as prosody, internal state, or perceived external characteristics of the conversation itself.

7. CONCLUSIONS

We have proposed a simple framework, based on the well-understood n-gram formalism, for modeling the sequences of conversational, participant-attributed vocal activity states. The proposed single-port non-parametric model, like previous work in dialogue but unlike that in multi-party conversation, treats participant states as conditionally independent of one another, given the joint past. We have argued that this allows for the synthesis of turn-taking patterns in a bottom-up fashion, as a function of the assumed impact of the degree of interlocutor overlap on one's own productions. The model was shown to lead to perplexity reductions which are at least as large as those of its multi-port counterpart. In addition, its performance was explored under extensions of the conditioning history and manipulation of its future-instant horizon. The techniques, we have argued, are likely to inform the design and behavior of spoken dialogue systems.

8. ACKNOWLEDGMENTS

This work was supported in part by Riksbankens Jubileumsfond (RJ), under contract P09-0064:1-E, *Samtalets Prosodi (Prosody in Conversation)*.

9. REFERENCES

[1] Paul T. Brady, "A model for generating on-off speech patterns in two-way conversation," *The Bell System Technical Journal*, vol. 48, no. 9, pp. 2445–2472, September 1969.

[2] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson, "A simplest semantics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, December 1974.

[3] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson, "Towards human-like spoken dialogue systems," *Speech Communication*, vol. 50, no. 8–9, pp. 630–645, August 2008.

[4] Elizabeth Shriberg, Andreas Stolcke, and Don Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. EUROSPEECH*, Aalborg, Denmark, September 2001, vol. 2, pp. 1359–1362.

[5] Joseph Jaffe, Stanley Feldstein, and Louis Cassotta, "Markovian models of dialogic time patterns," *Nature*, vol. 216, pp. 93–94, October 1967.

[6] Joseph Jaffe and Stanely Feldstein, *Rhythms of Dialogue*, Academic Press, New York NY, USA, 1970.

[7] Antoine Raux and Maxine Eskenazi, "Finite state turn taking model for spoken dialog systems," in *Proc. HLT-NAACL*, Boulder CO, USA, June 2009, pp. 629–637.

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

[9] 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.

[10] 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.

[11] E. Chapple, "The Interaction Chronograph: Its evolution and present application," *Personnel*, vol. 25, no. 4, pp. 295–307, January 1949.

[12] Emanuel A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, no. 1, pp. 1–63, March 2000.

[13] Gabriel Skantze and Anna Hjalmarsson, "Towards incremental speech generation in dialogue systems," in *Proc. SIGdial*, Tokyo, Japan, September 2010, pp. 1–8.