INTERSPEECH 2014

On the conversant-specificity of stochastic turn-taking models

Kornel Laskowski

Carnegie Mellon University, Pittsburgh PA, USA
Voci Technologies Inc., Pittsburgh PA, USA
kornel@cs.cmu.edu

Abstract

Stochastic turn-taking models provide stationary estimates of the probability of a conversant's incipient speech activity, given their own and their interlocutors' recent speech activity. Existing research suggests that such models may be conversant-specific, and even conversant-discriminative. The present work establishes this explicitly. It is shown that: (1) the conditioning context can be relaxed to exploit speech activity which need not be attributed to specific interlocutors; (2) the same duration of context can yield better results with a more statistically sound framework; and (3) results further improve asymptotically with the consideration of longer conditioning histories. The findings indicate that inter-conversant variability is a major contributor of variability across stochastic turn-taking models.

Index Terms: stochastic turn-taking, speech activity, speaker discrimination, speaking style

1. Introduction

Stochastic turn-taking (STT) models [1] are models of incipient, participant-attributed, binary-valued speech activity, conditioned on the recent conversational past. In their simplest incarnation, the conditioning history is limited to binary-valued speech activity. In that setting, they are most conveniently thought of as predictors of the *chronograms* [2] of conversations, causally from left to right; an example of such a chronogram is shown in Figure 1.

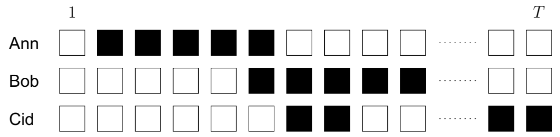

Figure 1: A vocal interaction record [3], or speech/non-speech chronogram [2], for a three-participant conversation whose duration is T 100-ms frames. Time shown from left to right, with frames numbered 1 through T; ■ and □ represent speech and non-speech, respectively.

Since their inception [4, 5] for two-party dialogue, bigram-based time-independent STT models have been extended to handle more than two conversants [3, 6], to condition predictions on long history durations [7], and to incrementally adapt to any time-dependent vagaries of conversations [8]. It is currently believed that STT model predictions are robust across participants and types of conversations, although not necessarily across corpus types and/or speech activity annotation or detection methodologies. What is less well understood is what accounts for the *variability* observed across STT models.

The current article is an effort to explain some of that variability, by asking the question

Q1. Does the identity of conversants in the training material affect model function?.

An affirmative answer would imply that the amount of intra-conversant variability is lower than the amount of inter-conversant variability, and vice versa. Work on time-dependent STT models [8] showed that individual test conversations deviate from model expectations, and that the deviation gap can be closed by incrementally adapting the model to the test conversation over time. This suggests the existence of inter- and/or intra- conversant variability, but does not evaluate their relative magnitudes.

To provide an answer to *Q1*, the current article applies time-independent, participant-specific models to chronograms of test conversations in which some of those same participants took part. It adopts the speaker attribution and speaker detection frameworks of [9], which also modeled chronograms but differently. It achieves error rates on completely held-out conversations which are 45.0%rel and 36.4%rel lower, on the attribution and detection tasks respectively, than reported in [9]. This result meets a more stringent criterion than necessary to answer *Q1* in the affirmative. It permits concluding that STT models exhibit considerably more inter-speaker that intra-speaker variability.

2. Data

As conversations, the experiments use 67 naturally-occurring, spontaneous-speech meetings of the ICSI Meeting Corpus [10], of type `Bed`, `Bmr`, and `Bro`. 33 meetings comprise the TRAINSET, while the DEVSET and EVALSET consist of 18 and 16 meetings, respectively. A chronogram of the type shown in Figure 1 is produced using the forced-alignment-mediated start and end times of lexical tokens transcribed manually from the close-talk channel of each participant. These start and end time are provided in the ICSI MRDA Corpus [11].

The number K of conversants is specific to each conversation, varying between 3 and 9. 14 of the participants in the 67 meetings each took part in a sufficient number of meetings to warrant training a conversant-specific model. The remainder of the participants became instances of the UNK speaker, for which a single separate model was trained in each experiment.

3. Baseline

The baseline system in the current article is taken directly from [9], where a speaker attribution task and a speaker detection task were treated separately. For the attribution task, the participants to a test conversation were assumed to be known in advance (also permitting the error-free inference of the type u

14 – 18 September 2014, Singapore

of the test conversation, with $u \in \{\texttt{Bed}, \texttt{Bmr}, \texttt{Bro}\}$). This required only that their optimal ordering, with respect to the rows of the test-conversation chronogram, be inferred. The inference was achieved using a *behavior model* of the group as a whole.

In the speaker attribution task, the identities of the participants to a test conversation of K participants were assumed not known a priori. Therefore, a group of K identities had to first be drawn from a population of putative conversants, using a *membership model*. The behavior model from the speaker attribution task was then applied to find the optimal permutation of each selected group.

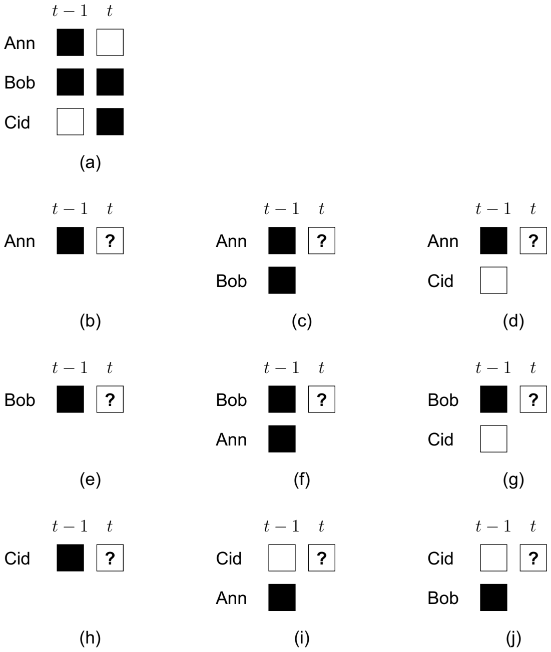

Figure 2: An instant t of a chronogram, in (a). The single event conditioned on the one-participant context for "Ann" is shown in (b), with the two events conditioned on the two alternative two-participant contexts for "Ann" are shown in (c). Panels (e) through (j) depict the situation for "Bob" and "Cid". It is noteworthy that in the baseline, the number of modeled events for each participant depends on the number of other participants.

Briefly, the behavior model for both tasks modeled participants in the context of their interaction with each of their interlocutors, taken one at a time. Instant t in the ficticious three-party conversation chronogram shown in Figure 2(a), for each of its three participants, yields one one-participant event (panels (b), (e), and (h)) and two two-partipants events (panels (c), (d), (f), (g), (i), and (j)). The features used by the baseline system are the probabilities of such events, modeled using Gaussian distributions. The overall training procedure for the task at hand is shown in Algorithm 1. Testing is conducted as in Algorithm 2. Both algorithms are shorthand for much more complete descriptions in [9].

4. Methods

Applying STT models to the tasks of speaker attribution and speaker detection permits replacing Algorithm 1 with Algorihm 3 and Algorithm 2 with Algorithm 4. This section describes the details in and differences among these algorithms; care has been taken to make the transition in small deltas, permitting an evaluation of the replacement of individual aspects.

Algorithm 1: Training in the baseline system

```
for each conversation $\mathcal{M}$ in TRAINSET do
    $K$ = number of participants in $\mathcal{M}$.
    Form $K$-row chronogram $\mathcal{C}$.
    Compute $K$ prior probabilities $V$ from $\mathcal{C}$.
    Compute $2^K \times 2^K$ transition probabilities $A$ from $\mathcal{C}$.
    Infer Ising model $\mathcal{I}$ from $A$.
    Accum $K$ one-participant features from $V$ and $\mathcal{I}$ into $\mathcal{H}_k$.
    Accum $K \times (K-1)$ two-participant features from $\mathcal{I}$ into $\mathcal{H}_{kk'}$.
for each of $N$ participants in population $\mathcal{P}$ do
    Build Gaussian $\mathcal{H}_k$.
    for each of $(N-1)$ interlocutors do
        Build Gaussian $\mathcal{H}_{kk'}$.
```

Algorithm 2: Testing in the baseline system

```
$K$ = number of participants in test conversation.
Form $K$-row chronogram $\mathcal{C}$.
Compute $K$ prior probabilities $V$ from $\mathcal{C}$.
Compute $2^K \times 2^K$ transition probabilities $A$ from $\mathcal{C}$.
Infer Ising model $\mathcal{I}$ from $A$.
Form $K$-row chronogram $\mathcal{C}$.
Compute $K$ prior probabilities $V$ from $\mathcal{C}$.
Compute $2^K \times 2^K$ transition probabilities $A$ from $\mathcal{C}$.
Infer Ising model $\mathcal{I}$ from $A$.
Compute $K$ one-participant features from $V$ and $\mathcal{I}$.
Compute $K \times (K-1)$ two-participant features from $\mathcal{I}$.
for each possible group of $K$ in $N$ do
    for each possible permutation of $K$ in $K$ do
        Score one-participant features using $\mathcal{H}_k$.
        Score two-participant features using $\mathcal{H}_{kk'}$.
        Compute joint likelihood.
Pick group and permutation which maximizes joint likelihood.
```

4.1. Disattributing interlocutor context

An important benefit of STT models is that they permit easy extension to arbitrarily long conditioning histories; by contrast, the baseline system characterized participants in terms of chronograms chopped up into intervals of two 100-ms frames. To make this benefit available to the task at hand, the baseline models have to be structurally simplified. In this article, that simplification comes from modeling participants in the context of a generic interlocutor, rather than in the context of each specific, named interlocutor. Panels (a), (d) and (g) in Figure 3 depict the creation of a 2-row chronogram for each row of the chronogram in Figure 2(a). The second row in each 2-row chronogram, which represents a virtual generic interlocutor labeled "oth" in Figure 3, contains the exclusive-OR of the speech activity states of the first-row-participant's actual interlocutors. This manipulation, shown to be useful in turn-taking modeling [7], yields only K one-participant events (panels (b), (e), and

(h) in Figure 3) and K two-participant events (panels (c), (f), and (i)).

Algorithm 3: Training in the STT system

```
for each conversation $\mathcal{M}$ in TRAINSET do
    $K$ = number of participants in $\mathcal{M}$.
    Form $K$-row chronogram $\mathcal{C}$.
    Form $K$ 2-row chronograms $\mathcal{C}_k$ from $\mathcal{C}$.
    for each $k \in [1, K]$ do
        Accum conditional counts of first row of $\mathcal{C}_k$ into
        $\mathcal{H}_k$.
for each of $N$ participants in population $\mathcal{P}$ do
    Build $N$-gram $\mathcal{H}_k$.
```

Algorithm 4: Testing in the STT system

```
$K$ = number of participants in test conversation.
Form $K$-row chronogram $\mathcal{C}$.
Form $K$ 2-row chronograms $\mathcal{C}_k$.
for each possible group of $K$ in $N$ do
    for each possible permutation of $K$ in $K$ do
        Score test $\mathcal{C}_k$ using $\mathcal{H}_k$.
        Accumulate joint likelihood.
Pick group and permutation which maximizes joint
likelihood.
```

A consequence of a transition to 2-row chronograms is that the Ising model formalism [12] employed in [9] is no longer necessary for the estimation of interlocutor-row-conditioned likelihoods.

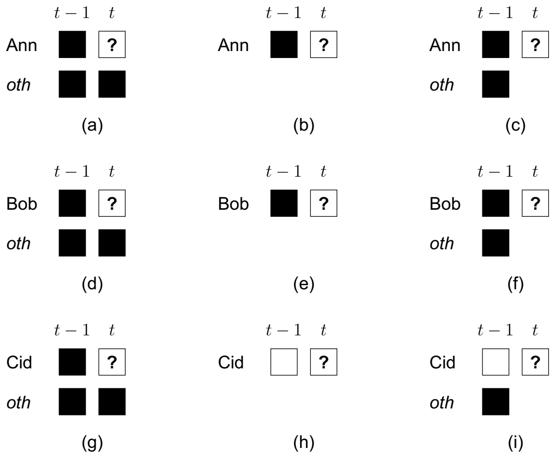

Figure 3: The same instant as in Figure 2(a); for each participant, the 3-row chronogram has been collapsed into a two-row chronogram (panels (b), (d), and (g)) whose second row depicts the exclusive-OR of that participant's interlocutors' speech activity states. The number of events modeled for each participant is *not* a function of the number of participants.

4.2. Direct first-order model evaluation

Provided that a K-row multi-party chronogram can be conceptually replaced with K 2-row "dialogue" chronograms, any conversant can be modeled using a structure which is independent of the number and identities of their interlocutors. STT models have precisely this characteristic; since they provide the likelihood of a chronogram row, there is no need on an intermediate modeling structure.

4.3. Extension to longer conditioning histories

Finally, provided that 2-row chronograms can be successfully modeled using first-order STT models, they can also be modeled using larger-order STT models. In the current article, contexts as long as 1 second, consisting of 10 100-ms frames, are considered. A conditionally independent STT N-gram model with this context is a 21-gram.

5. Experiments

5.1. Known conversation type and known particant group

Knowing the conversation type permits selecting training material from conversations of that type alone, while knowing the identities of the conversants yields the speaker attribution task. For a test conversation of K participants, the search space consists of $K!$ equi-probable orderings. As in [9], these are evaluated exhaustively.

Figure 4 shows the DEVSET classification error rate of the baseline [9] (as described in Section 3); it also shows the performance of that baseline with only maximum likelihood model estimation, e.g. no smoothing ("noS"). Smoothing was removed to more easily assess subsequent developments; it can be seen however that the smoothing proposed in [9] had reduced the baseline error rate by almost 10%abs.

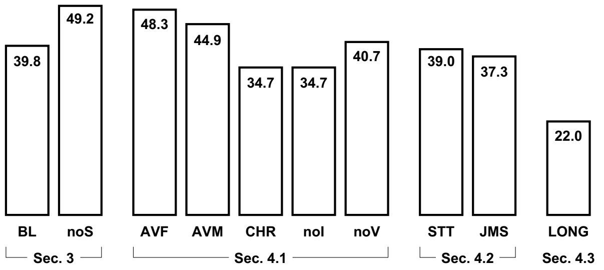

Figure 4: Classification error rates (%) for DEVSET on the exhaustive-search speaker attribution task, at various stages during development. "BL": baseline, "noS": BL without smoothing, "AVF": noS with averaging interlocutor features during testing, "AVM": AVF with averaging interlocutor features during training, "CHR": AVM with features drawn from 2-row chronograms, "noI": CHR with features estimated not using the Ising model formalism, "noV": noI with unigram probability-of-speaking features removed, "STT": bigram stochastic turn taking models applied directly to chronograms, "JMS": STT with Jelinek-Mercer smoothing, "LONG": JMS with greater-that-first-order models.

The elimination of interlocutor identities, in the features for individual participants, is shown in columns 3 through 6 of Figure 4. The averaging over interlocutors of two-participant features, during testing ("AVF") and also training ("AVM"), yields modest reductions over "noS" of 4.3%abs. Drawing two-participant features from two-row chronograms, representing each actual participant in the context of a virtual interlocutor (representing the inclusive-OR of the speech activity of actual interlocutors) reduces error rates by more than 10%abs. This permits for the elimination of Ising model estimation in the processing pipeline of [9], in system “noI”.

Finally, column 7 in Figure 4 indicates that excluding the single unigram feature per participant, namely the probability of speaking, is deleterious; it leads to a 6.0%abs increase in DevSet error rates. The reason for this modification is to directly compare with the stochastic turn-taking model approach, in which single models are N-grams of a fixed order. First-order N-gram STT models (column 8), without Jelinek-Mercer interpolation [13], are seen to be only 1.7%abs better. However, interpolation (column 9) reduces error rates by an additional 1.7%abs. Extending the conditioning history to 10 frames of context (column 10) reduces the error rate by a further 15.3%abs. The 10th column, denoted “LONG”, represents a 44.7%rel reduction of DevSet error from the baseline.

5.2. Known type but unknown group

When the identifies in the conversation group are unknown, they must first be drawn from a population of N putative participants. This requires drawing of a combination of K items from N, prior to considering the $K!$ permutations of each draw, thus making exhaustive search generally intractable. For this reason, a greedy algorithm is proposed for the speaker detection task. At each iteration: (1) the lowest negative log-likelihood in the $\{$ test conversation participants $\} \times \{$ available models $\}$ search matrix is selected; and (2) the model is excluded from subsequent iterations unless it is the Unk participant model.

This greedy algorithm was first applied to the speaker detection task of the previous subsection, to estimate the cost of *not* performing an exhaustive search. Figure 5 shows the exhaustive-search baseline (“BL”) and the exhaustive-search STT-based system (“EXH”) error rates in the first two columns; it also shows the performance of the greedy-search STT-based system (“GR”). As can be seen, “GR” incurs unacceptably higher error rates. It appears that this is due largely to the fact that some test conversation participants are easier to predict than others, and these are picked off first by the algorithm. To address this problem, a “GRZ” variant was proposed which first Z-normalizes the search matrix test-participant columns. On DevSet, “GRZ” outperforms exhaustive search in the unnormalized search matrix.

“GRZ” was then applied to the speaker detection task. When the test conversation type is known, the population of training participants is smaller than when the type is unknown. Column 5 in Figure 5 shows that error rates are approximately 7%abs higher than for the speaker attribution task. ([9] did not contain a comparable evaluation.)

5.3. Unknown type and unknown group

When also the conversation type is unknown, all training conversation types must be considered when a new test conversation is analyzed. “GRZ” for this case, shown in column seven of Figure 5, is higher than in the preceding subsection. However, it is already an improvement over the baseline (“BL”, column six). Results can be improved further by requiring that the system implicitly identify the conversation type, via an argmax versus a sum in the Bayesian computations leading to the search matrix. The reason that the resulting “GRZT” system is better than “GRZ” is that it forbids combinations of participants which attended different types of training conversations. As can be seen, “GRZT” reduces the DevSet error rate from 61.0% in [9] to 32.2%, by 47.2%rel.

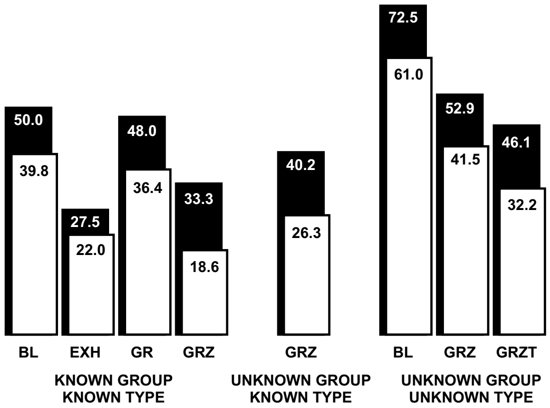

Figure 5: Classification error rates (%) for DevSet and EvalSet in white and black, respectively. “BL”: baseline, “EXH”: exhaustive search, “GR”: greedy search, “GRZ”: greedy search with Z-normalization, “GRZT”: greedy search with Z-normalization and implicit conversation type selection.

6. Discussion

6.1. Generalization

As can be seen in Figure 5, the performance trends observed on DevSet are approximately the same as those on the completely held-out EvalSet. On the exhaustive-search speaker attribution task, the EvalSet error rate is reduced from 50.0% to 27.5%, by 45.0%rel; the reduction on DevSet was 44.7%rel. On the greedy-search speaker detection task, the EvalSet error rate is reduced from 72.5% to 46.1%, by 36.4%rel. This is considerable, although smaller than the 47.2%rel reduction for DevSet. The results suggest that the improvements reported in this work broadly generalize to unseen data.

6.2. Impact

Although discriminating among conversants was only a means and not a goal of the current article, it is conceivable that STT models will in the future be used for discrimination, particularly in privacy-sensitive or otherwise ablated contexts. The current work also has the potential to help in the inference of prototypical conversational behavior styles.

7. Conclusions

This article hypothesized that the variability observed in STT models trained on speech/non-speech chronograms is due in large part to inter-conversant variability. Conversant classification experiments were conducted to test this hypothesis. On the attribution and detection tasks, error rates for completely held-out sets of conversations were shown to be 45.0%rel and 36.4%rel lower than for previously published systems, which also relied exclusively on chronograms. This indicates that intra-speaker variability is smaller than inter-speaker variability, in the STT model formalism, and entails that conversants exhibit approximately stationary STT statistics across conversations. The findings offer promosing avenues for the description of conversational turn-taking style.

8. References

[9] Laskowski, K. and Schultz, T., “Recovering participant identities in meetings from a probabilistic description of vocal interaction”, in *Proc .9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brisbane, Australia, pp. 82-85, September 2008.

[7] Laskowski, K., Edlund, J. and Heldner, M., “A single-port non-parametric model of turn-taking in multi-party conversation”, in *Proc. 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Praha, Czech Republic, pp. 5600-5603, May 2011.

[8] Laskowski, K, Edlund, J. and Heldner, M., “Incremental learning and forgetting in stochastic turn-taking models”, in *Proc. 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Firenze, Italy, pp. 2065-2068, August 2011.

[12] Glauber, R., “Time-dependent statistics of the Ising model”, in *Journal of Mathematical Physics*, **4**(2):294-307, 1963.

[5] Laskowski, K., Ostendorf, M. and Schultz, T., “Modeling vocal interaction for text-independent classification of conversation type”, in *Proc. 8th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, Antwerpen, Belgium, pp. 1258-1261, September 2007.

[10] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. and Wooters, C., “The ICSI Meeting Corpus”, in *Proc. 28th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, China, pp. 364-367, April 2003.

[11] Shriberg, E., Dhillon, R., Bhagat, S., Ang, S. and Carvey, H., “The ICSI Meeting Recorder Dialog Act (MRDA) Corpus”, in *Proc. 5th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, Cambridge MA, USA, pp.97-100, April 2004.

[5] Brady, P., “A model for generating on-off speech patterns in two-way conversation”, in *The Bell System Technical Journal*, **48**(9):2445-2472, September 1969.

[4] Jaffe, J., Feldstein, S. and Cassotta, L., “Markovian models of dialogic time patterns”, in *Nature*, **216**:93-94, October 1967.

[2] Chapple, E., “The Interaction Chronograph: Its evolution and present application”, in *Personnel*, **25**(4):295-307, January 1949.

[13] “Interpolated estimation of Markov source parameters from sparse data”, in *Proc. Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands, 1980.

[1] Wilson, T., Wiemann, J. and Zimmerman, D., “Models of turn-taking in conversational interaction”, in *Journal of Language and Social Psychology*, **3**(3):159-183, 1984.

[3] Dabbs Jr., J. and Ruback, R., “Dimensions of group process: Amount and structure of vocal interaction”, in *Advances in Experimental Social Psychology*, **20**:123-169, 1987.

[6] Laskowski, K., “Modeling norms of turn-taking in multi-party conversation”, in *Proc. 48th Annual Meeting of Association for Computational Linguistics (ACL)*, Uppsala, Sweden, pp.999-1008, July 2010.