
SOCIAL CORRELATES OF TURN-TAKING BEHAVIOR

*John Grothendieck**, *Allen Gorin*[†], and *Nash Borges*[‡]

BBN Technologies*
U.S. Department of Defense^{†‡}
Johns Hopkins University[‡]

ABSTRACT

The goal of this research is to infer traits about groups of people from their turn-taking behavior in natural conversation. These traits are latent attributes in a social network, whose relative frequencies we estimate from content-derived metadata. Our approach is to train statistical models of turn-taking behavior using automatic labels of speech activity, and measure the association of these models with socially correlated traits. We experimentally evaluate these ideas using the Switchboard-1 speech corpus, which provides speech content and metadata associated with each speaker, such as gender, age and education, as well as inferred social correlates such as willingness to participate and initiate. We show that population proportions of these socially correlated externals can be predicted with a root mean-squared error of approximately 0.1 across all mixture proportions.

Index Terms— turn-taking, social network analysis, speech detection

1. INTRODUCTION

The goal of this research is to infer traits about groups of people from their spoken conversations. There are many practical and theoretical motivations for this problem. Potential applications include detecting dissatisfied customers and predicting influence between speakers for social network analysis. Theoretical interest centers on placing such information within a broader context. Social network analysis captures structural information about human interactions [1]. A set of conversations amongst individuals can be organized into an attributed communication graph with speakers as vertices and dialogs as edges. The content of individual messages can be analyzed using automatic language processing. However, exploiting dependencies among content and other attributes of the communication graph is a relatively new research area. Context can improve language models as in [2]; here we consider the converse, using content models to improve estimation of unknown externals. A similar approach was taken by [3], although their corpora and content models differ from those used in this paper.

Many socially correlated traits are not directly observable from the words that are spoken, but are revealed from speaker interactions. These traits include attributes of a speaker, such as age, education, gender, and where they were raised. We refer to these as external metadata, Ξ^E . Other traits are attributes of how a speaker interacts with others, such as the frequency of communications. We denote such metadata, derived from a communication graph, as Ξ^G . Meta-

data can also be derived from the content of a communication, which we refer to as meta-content and denote Ξ^C .

Conversation consists of related streams of information. Multi-media analysis of group meetings has led to the notion of dialog as a group activity, with individual actions impacting a collective state [4]. In this work, we train models of the stochastic process of turn taking from observed speech activity states in two-person dialogs. This ignores linguistic content, but even structural dialog acts can aid speech recognition [5]. Activity can be modeled over small or large portions of dialog [6]; we consider actions corresponding to such notions as sentence fragment or speaker turn.

How a conversation proceeds can be more important than what is said [7]. We wish to relate turn-taking behavior to socially correlated externals. Existing research suggests that social role, age, gender, education, and background culture all impact turn-taking behavior [3] [8]. (So may dialog type [9], topic [10], or familiarity among speakers [11].) Relatively simple audio features can be used to model turn-taking behavior. Systematic differences emerge in models trained from different sets of speakers or dialogs.

The Switchboard-1 corpus includes demographic information about the participants in addition to the dialogs. While the speakers and dialogs in Switchboard do not comprise a natural social network, speakers' levels of participation and initiation vary. Thus metadata Ξ^E , Ξ^G , and Ξ^C are all present, and social correlates of turn-taking behavior can be estimated and exploited.

We will show how to train turn-taking models from Switchboard, using dialog states inferred from speech activity detection (SAD). This allows clustering of speakers based on similarity of their turn-taking behavior. We observe statistical dependence of cluster membership with other speaker traits on the training data, i.e. various traits are correlated with a speaker's derived turn-taking "style." Experimentally we sample test data using different proportions of turn-taking style. Over the test data, we estimate proportions of the correlates of turn-taking style from observed turn-taking behavior.

The rest of the paper is organized as follows. A brief summary of the Switchboard-1 communication graph is provided in Section 2. A model of speaker turn-taking behavior is presented in Section 3. Section 4 describes the experimental results and data analysis, and Section 5 presents concluding remarks.

2. SWITCHBOARD-1 COMMUNICATION GRAPH

The Switchboard-1 corpus [12] has been used to evaluate many problems in language processing. In this work, Switchboard provides a convenient test-bed for a proof-of-concept experiment. It has extensive, publicly available annotations, including manual word-level transcriptions of all dialogs, with time alignments for both sides. Content consists of 2438 dialogs, recorded in two-channel audio

*jgrothen@bbn.com

†a.gorin@ieee.org

‡nashborges@jhu.edu

files. Each dialog is a variable-duration topical conversation between exactly two of the 520 participants.

The set of spoken dialogs between speakers can be considered as a communication graph. A communication (dialog) is an attributed edge connecting the speakers. Dialog attributes include date, time and topic. This work focuses on speaker attributes. Speaker attributes such as willingness to participate correspond to the degree of a vertex (speaker) in the graph, as does willingness to initiate a dialog (the out-degree of a vertex). Some speakers always initiate dialogs, others never do so. Such graph-derived metadata is extracted from the structure in Figure 1.

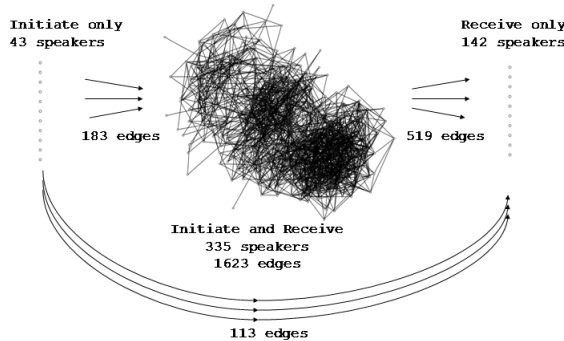


Fig. 1. Switchboard-1 communication graph.

A few demographic fields were recorded for the participants. These included accent, age, education level, and gender. Additional attributes can be derived from content; in particular the parameters of a model $D(t)$ of a particular speaker’s turn-taking behavior can be trained on dialog activity from all adjacent edges. Clustering speakers with similar turn-taking behavior allows the categorical attribute of cluster membership to serve as shorthand for speaker turn-taking styles.

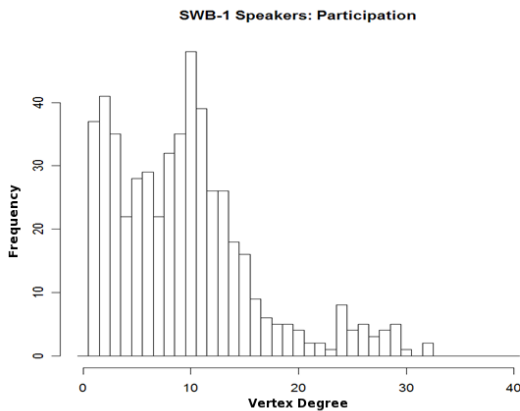


Fig. 2. SWB-1 speaker participation.

Due to the data collection protocol [13], two participants, typically strangers, have at most one mutual conversation. Vertex degree depended on voluntary participation in dialogs, shown in Figure 2,

with an average participation of roughly 9 dialogs per speaker. The data collection methodology was asymmetric in that dialogs were initiated by a speaker — Speaker 1 called the robot operator, which found some other study member willing to participate at that moment. Thus in-degrees differ systematically from out-degrees. “Participation” and “initiation” levels are speaker attributes derived from the communication graph.

The Switchboard dialogs reside in the attributed graph G as sketched in Figure 1. The corpus permits study of dependencies among Ξ^C , Ξ^E , and Ξ^G , in particular between turn-taking behavior and other socially correlated speaker traits.

It remains to provide details of training the turn-taking model from Switchboard. For SAD, we use the energy-based *xtalk* tool from MIT Lincoln Laboratory [14]. This produces a stream of 10 ms frame labels of speech activity detection for each dialog side, which we smooth via a 10-frame window. One-sided segmentation proceeds as follows: inactive regions of at least 2 seconds length are marked as I, others as A, and adjacent states of the same type combined to create a one-sided alternating activity sequence, e.g. IAI...AI. We combine both sides to create a dialog activity state sequence $S(t)$ as in Figure 3. The $S(t)$ are used to train models of dialog turn-taking $D(t)$.

Side 1: $S_1(t)$	I	A	I	A
Side 2: $S_2(t)$	A	I	A	I
Dialog State: $s(t)$	IA	II	AI	AA

Fig. 3. Dialog state from multiple sides.

Since Switchboard-1 has been thoroughly transcribed, we can verify our results using the manual transcriptions without the errors inherent in automatic labeling. Switchboard audio clips frequently include distinct crosstalk from the other speaker, which is quite properly detected by a sensitive SAD system. This raises a challenge in detecting which speaker is active at which time (see for example [15] for a detailed treatment). Our SAD results depart from ground-truth dialog states with a frame error of approximately 16%.

3. MODELS OF SPEAKER TURN-TAKING BEHAVIOR

The actions of one participant in a conversation is a sequence in time, with regions of both speech and silence. Denoting activity by “A” and inactivity by “I”, a single speaker produces a sequence of these two states. For a two-sided dialog, this suggests two aligned sequences with speech on the part of one speaker slotting neatly into a listening silence from the other side. The reality is more complex; silent regions arise when speakers pause to offer the floor or to breathe, while interjections or miscues result in audio from both speakers simultaneously. Audio records of a two-sided dialog can be partitioned into four states: AA, AI, IA, and II, as shown in Figure 3.

Consider a semi-Markov process, with history of length $k \geq 0$ leading to state transition probabilities $P(X_t|X_{t-k}, \dots, X_{t-1})$ and state durations $f(X_t|X_{t-k}, \dots, X_t)$. We prohibit self-transitions. State durations are modeled via Gamma distributions, which allow a qualitatively reasonable fit to the empirical distributions, and seem theoretically plausible as sums of independent exponential durations.

Given a dialog state sequence $S(t)$, state counts and durations (conditioned on history) allow estimation of transition probabilities and duration distribution parameters. We use maximum-likelihood estimates of model parameters.

Shifting consideration from dialogs to speakers raises a few issues. While dialogs may divide into “Side 1” versus “Side 2”, speakers do not. A participant in a two-sided dialog perceives four activity states, but the natural interpretation differs: AA (both speakers active), AI (self active), IA (other active), and II (neither active). While this combines all “other” people into one class, modeling relationships between individuals is beyond the scope of this work. Note that this creates two “perspectives” for a two-sided dialog, with AI and IA exchanged depending on “self.”

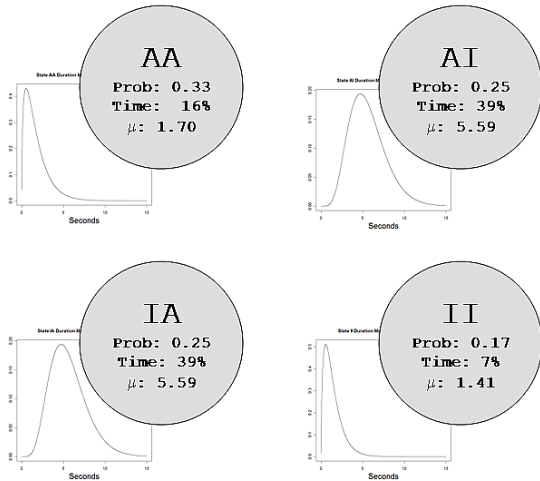


Fig. 4. Unigram turn-taking model, trained from all *xtalk* turn-level speaker activity sequences.

We concatenate the $S(t)$ of all dialogs including speaker V (from his or her perspective) to generate speaker perspective sequences $S_V^P(t)$. Combining all $S_V^P(t)$ provides the data for a general content model as shown in Figure 4. While state probabilities are natural model parameters, relative time spent in a state seems more useful for understanding (and is equivalent given average state durations). AI and IA are equivalent in global models since both perspectives of every dialog are used to train: every state IA appears elsewhere as an AI. Note this is not the case for subsets of the corpus; in a content model trained from $S_V^P(t)$ the speaker-perspective sequence for speaker V , states AI (V speaking) and IA (other speaking to V) can have very different properties.

It is straightforward to estimate ML turn-taking model parameters from a dialog activity sequence. We apply a standard divisive clustering algorithm [16]:

- For each speaker V , create speaker-perspective training sequence $S_V^P(t)$.
- Train model of overall $D_G(t)$ from $\cup S_V^P(t)$.
- Randomly partition speakers into sets C_0 and C_1 .
- Iterate until convergence:
 - Train cluster models $D_i(t)$ from sequences $S_j^P(t)$, $j \in C_i$.
 - Reassign each speaker j into the C_i for which $S_j^P(t)$ has minimum cost (here negative log-likelihood) under the model of $D_i(t)$

We henceforth denote members of C_i as having turn-taking style i .

The model parameters resulting from a particular initial partition are shown in Table 1. Here $|C_0| = 306$ and $|C_1| = 214$. Style 0 is characterized by fewer and shorter states AA, more and longer II, slightly longer AI, and longer IA than style 1. A typical Switchboard dialog provides enough evidence to discriminate between styles; 5-minute sequences generated according to models $D_0(t)$ and $D_1(t)$ can be classified with approximately 0.2% EER. Turn-taking style shows significant dependence with other traits, as shown in Table 2.

State	Time ₀	μ_0	σ_0	Time ₁	μ_1	σ_1
AA	0.09	1.16	1.26	0.22	2.07	2.48
AI	0.40	5.81	5.40	0.38	5.38	4.30
IA	0.42	6.30	6.02	0.35	4.94	3.73
II	0.09	1.50	1.25	0.04	1.27	1.12

Table 1. Model parameters for speaker turn-taking styles: relative time in state, and mean and standard deviation of durations Gamma(α, β).

Trait	% C_0	% C_1	P-value
Gender(Female)	31	67	2e-17
Accent(South Midland)	23	41	5e-06
Initiation(Never)	32	21	0.003
Education(Graduate)	36	28	0.02

Table 2. Selected trait proportions and (unadjusted) p-values. Thus 31% of style 0 speakers are female, as opposed to 67% of style 1 — gender is not homogeneous across turn-taking styles.

That gender relates to conversational style is well-known from linguistics [17]. Thus the correlation of turn-taking behavior with gender is not surprising. However, the nature of other observed correlations is less evident. Insofar as these capture general relationships (rather than being artifacts of the corpus), these provide a means for characterizing speaker traits on a novel data set via robust natural language processing.

4. EXPERIMENTS

4.1. Prediction of Externals from Turn-Taking Behavior

Cluster-conditional trait distributions can be used to estimate test set relative frequencies. Let V_i be the estimated (multinomial) distribution for some trait in cluster C_i . Let λ be the estimated cluster proportions within the test set. Thus we have mixture model

$$V_{pred} = \lambda V_0 + (1 - \lambda) V_1 \quad (1)$$

We use λ the relative frequency of hard cluster membership assignments of test observations.

Cluster properties (the association of other traits with turn-taking behavior) may prove more portable across data sets than mixture proportions. We partition Switchboard to investigate performance on matched data, but with different proportions of turn-taking behavior. The set of speaker sequences $\{S_i^P(t)\}$ are randomly divided into training and test sets (here sizes 260:260). The training set is used to train cluster models of $D_0(t)$ and $D_1(t)$ as before. Test speakers are assigned to clusters based on minimum sequence cost under the model of $D_i(t)$. Evaluation sets (here of size 100) are chosen at random from the test members in each cluster to cover a desired

range of $\lambda \in [0, 1]$ (here 5% increments). We measure predictive performance via squared error loss, $SE = (V_{pred} - V_{obs})^T (V_{pred} - V_{obs})$. This investigates a range of evaluation sets, but has fixed training set. We can estimate $RMSE = \left(\frac{1}{T} \sum_{j=1}^T SE_j\right)^{1/2}$ using squared error over a batch of such experiments (here 200).

Results for selected traits are shown in Figure 5. Note that for consistency with earlier notation, training clusters are relabeled — the characteristics of styles 0 and 1 are analogous to those in Section 3 perhaps 99% of the time. RMSE is as expected from the variance due to limited training and evaluation set sizes, with no evident bias for different mixture proportions. The speaker traits displayed are all categorical, but differ in number of possible values and entropy; this explains the different RMSE levels. Strength of correlation with turn-taking behavior also varies, but plotted curves are nearly flat since Equation 1 incorporates mixture parameter λ . For comparison, RMSE from the trait distributions observed on the training data are also presented; models of turn-taking style improve estimates of gender and accent on the test data.

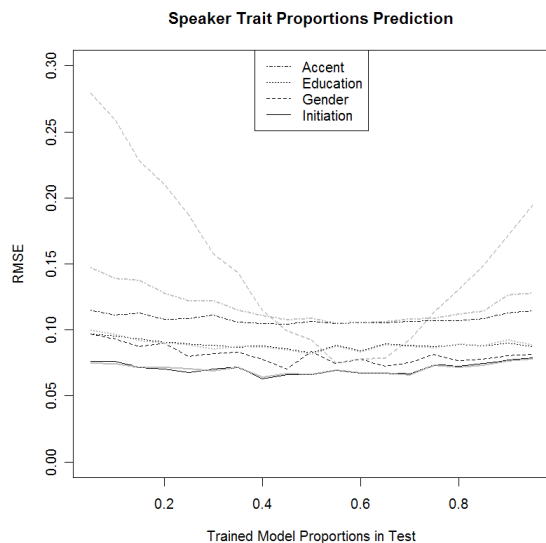


Fig. 5. Test distribution RMSE for selected speaker traits over a range of C_0 proportions in the test data, from Equation 1 (black) over all training distribution (gray).

5. CONCLUSIONS

This paper has addressed the problem of estimating the proportion of social correlates from content-derived turn-taking behavior. Some interesting points emerge from our methodology: a simple stochastic model captures certain aspects of turn-taking behavior, systematically different turn-taking styles emerge from clustering via these models, and turn-taking style is related to other attributes. Our models can be trained from automatic speech activity marks, so inference about traits such as education is possible without any manual annotations.

These results are one step within a general program of using known attributes of a communications graph to estimate desired unknowns. While speakers and dialogs are fundamentally different objects, both lie within the communication graph structure. Thus un-

derstanding of low-level audio features is related to understanding of such generalities as human behavior.

6. REFERENCES

- [1] C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park, “Scan Statistics on Enron Graphs,” *Computational and Mathematical Organization Theory*, vol. 11, no. 3, pp. 229–247, 2005.
- [2] A. McCallum, A. Corrada-Emmanuel, and X. Wang, “Topic and Role Discovery in Social Networks with Experiments with Enron and Academic Email,” *Journal of Artificial Intelligence Research*, vol. 30, pp. 249–272, 2007.
- [3] K. Laskowski, M. Ostendorf, and T. Schultz, “Modeling Vocal Interaction for Text-Independent Participant Characterization in Multi-Party Conversation,” in *Proc. SIGdial*, 2008, pp. 148–155.
- [4] D. Zhang, S. Bengio, D. Gatica-Perez, and D. Roy, “Learning Influence Among Interacting Markov Chains,” in *NIPS 2005*, 2005.
- [5] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, “Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech,” *Computational Linguistics*, vol. 26, no. 3, 2000.
- [6] S. Basu, *Conversational Scene Analysis*, Ph.D. thesis, MIT, 2002.
- [7] A. Pentland, *Honest Signals*, MIT Press, 2008.
- [8] A. Berry, “Spanish and American turn-taking styles: A comparative study,” *Pragmatics and Language Learning, monograph series 5*, pp. 180–190, 1994.
- [9] L. ten Bosch, N. Oostdijk, and J.P. de Ruiter, “Durational Aspects of Turn-Taking in Spontaneous Face-to-Face and Telephone Dialogues,” in *TSD*, 2004, pp. 563–570.
- [10] K. Laskowski, M. Ostendorf, and T. Schultz, “Modeling Vocal Interaction for Text-Independent Classification of Conversation Type,” in *Proc. SIGdial*, 2007, pp. 194–201.
- [11] J. Jaffe, B. Beebe, S. Feldstein, C.L. Crown, and M.D. Jasnow, “Rhythms of Dialogue in Infancy: Coordinated Timing in Development,” *Monographs of the Society for Research in Child Development*, vol. 66, no. 2, pp. 1–132, 2001.
- [12] J. Godfrey and E. Holliman, “SWITCHBOARD-1 Release 2,” Linguistic Data Consortium, LDC97S62, 1997.
- [13] “SWITCHBOARD: A User’s Manual,” http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html.
- [14] D.A. Reynolds, *A Gaussian Mixture Modelling Approach to Text-Independent Speaker Identification*, Ph.D. thesis, Georgia Institute of Technology, August 1992.
- [15] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals, “Speech and Crosstalk Detection in Multichannel Audio,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.
- [16] P. Chou, “Optimal Partitioning for Classification and Regression Trees,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 340–354, 1991.
- [17] D. Tannen, Ed., *Gender and Conversational Interaction*, Oxford University Press, 1993.