A Framework for the Automatic Inference of Stochastic Turn-Taking Styles

Kornel Laskowski
Carnegie Mellon University, Pittsburgh PA, USA
Voci Technologies, Inc., Pittsburgh PA, USA

Abstract

Conversant-independent stochastic turn-taking (STT) models generally benefit from additional training data. However, conversants are patently not identical in turn-taking style: recent research has shown that conversant-specific models can be used to refractively detect some conversants in unseen conversations. The current work explores an unsupervised framework for studying turn-taking style variability. First, within a verification framework using an information-theoretic model distance, sides cluster by conversant more often than not. Second, multi-dimensional scaling onto low-dimensional subspaces appears capable of preserving distance. These observations suggest that, for many speakers, turn-taking style as characterized by time-independent STT models is a stable attribute, which may be correlated with other stable speaker attributes such as personality. The exploratory techniques presented stand to benefit speaker diarization technology, dialogue agent design, and automated psychological diagnosis.

1 Introduction

Turn-taking is an inherent characteristic of spoken conversation. Among models of turn-taking (Jaffe et al., 1967; Brady, 1969; Wilson et al., 1984; J. Dabbs and Ruback, 1987; Laskowski, 2010; Laskowski et al., 2011b), those labeled "stochastic turn-taking models" (Wilson et al., 1984) offer a particular advantage: they are independent of the meaning of just what a "turn" might be. This is felicitous, since researchers are in disagreement over the definition. Instead, stochastic turn-taking (STT) models provide a probability that a specific participant speaks at instant t, conditioned on what that participant and her interlocutors were doing at specific prior instants. Whether her speaking constitutes something that might be called a "turn" is not germane to the applicability of STT models.

In their most commonly studied form (Jaffe et al., 1967; Brady, 1969; Laskowski, 2010), STT models condition their estimates on a history that consists exclusively of binary speech/non-speech variables; the extension to more complex characterizations of the past have been studied (Laskowski, 2012) but comprise the minority. In this binary-feature mode of operation, STT models ablate from conversations the overwhelming majority of the overt information contained in them, including topic, choice of words, language spoken, intonation, stress, voice quality, and voice itself, leaving only speaker-attributed chronograms (Chapple, 1949) of binary-valued behavior. This is a strength particular to STT models: they are language-, topic-, and text- agnostic, and therefore stand to form a universal framework for comparison of conversational behavior, where other methods would need to be extended to cross language, topic, and speech usage boundaries.

Given the paucity of information contained in chronograms, however, it is surprising that they have been efficiently exploited in the supervised tasks of conversation-type inference, participant-role inference, social status inference, and even identity inference. The current article aims to extend understanding of STT models in an unsupervised way, by starting from a theoretically sound distance metric between models of individual, interlocutor-contextualized conversation sides. In the space induced by these distances, experiments and analyses are performed which aim to answer a fundamental question: *Do people behave self-consistently, across disparate longitudinal obser-*

Proceedings of the SIGDIAL 2016 Conference, pages 202–211,
Los Angeles, USA, 13-15 September 2016.

vations, in terms of their turn-taking preferences? (Self-consistency *within* conversations was studied indirectly in (Laskowski et al., 2011b).) To provide an answer, between-person scatter is compared to within-person scatter, and accounts are sought for both types of variability. The findings reveal that models of persons are in fact self-consistent on average, and that, therefore, both (1) the persons they model are self-consistent, and (2) the modeling framework presented here is capable of capturing that self-consistency, while simultaneously differentiating among persons. The work has important implications for social psychology, diarization technology, and dialogue system design.

2 Data

The data used in this work was drawn from the ICSI Meeting Corpus (Janin et al., 2003), which consists of 75 multi-party meetings involving naturally occurring, spontaneous speech. It has been claimed that the meetings would have taken place even if they were not being recorded.

DATASET as defined here is limited to all 29 of the `Bmr` meetings, i.e. those held by the group of 15 researchers working on the Meeting Recorder project at ICSI. Not all 15 persons participated in every meeting; each of the 29 meetings was attended by an average of 6.8 persons. The total number of conversation *sides* in DATASET is 197. The distribution of sides per participant is shown in Figure 1.

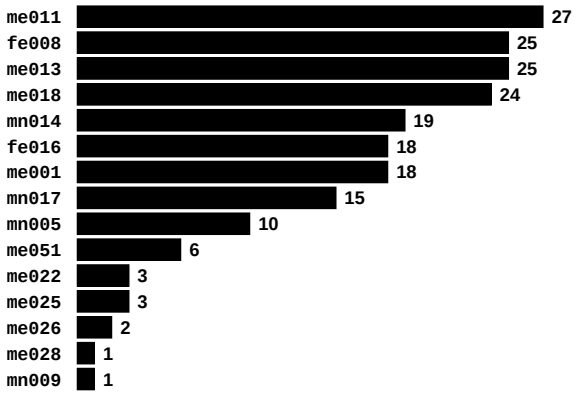


Figure 1: The number of sides in DATASET contributed by each of its 15 participants.

Each meeting in the ICSI Meeting Corpus contains an interval of time (at the beginning or end of the meeting) marked as `Digits`, used for microphone calibration. This interval was excluded for the current purposes, as it does not involve conversation. Each recording was left with between 22.8 and 74.5 minutes of data, with an average of 48.4 minutes.

3 Methodology

3.1 Chronograms

From each meeting $\mathcal{C}$ in DATASET, a speech/non-speech chronogram (Chapple, 1949) was constructed, designated by $\mathbf{Q}$. $\mathbf{Q}$ is a matrix whose entries are one of $\{\square, \blacksquare\}$, or equivalently $\{0, 1\}$, designating non-speech or speech respectively. Rows represent the K persons participating in the meeting, while columns represent 100-ms time frames covering its temporal support. The average $\mathbf{Q}$ in DATASET thus contained $K = 7$ rows and $T = 29$K columns.

The cell in row k and columnt t of every $\mathbf{Q}$ was populated, by a value of $\square$ or $\blacksquare$, by inspecting the forced alignments to the manually transcribed speech attributed to the kth speaker of the corresponding meeting. The transcriptions, attributions, and alignments had been made available by ICSI in (Shriberg et al., 2004). A frame increment of 100 ms was chosen as in (Laskowski et al., 2011b) and (Laskowski et al., 2011a); this is shorter than the average syllable duration, ensuring that no speech is missed, but longer than the frame step of the recognizer used by ICSI for the forced alignment. This makes the models developed in the current work robust to imprecision in word start and end times.

3.2 Stochastic Turn-Taking Models

The models used in the current work are probabilistic generative models that, given a chronogram $\mathbf{Q} \in \{\square, \blacksquare\}^{K \times T}$, provide the probability that its kth participant will speak during its tth frame. Participants are most commonly (Jaffe et al., 1967; Brady, 1969; Laskowski et al., 2011b) treated as conditionally independent (or "single-source" in the terminology of (Jaffe et al., 1967)); namely, the probability of speaking at frame t for participant k is independent of what the other $K - 1$ participants do at frame t, but it is conditioned on the joint K-participant history. The history duration, in number of most-recent contiguous frames, is denoted henceforth by τ.

In multi-party conversation, the number K of participants varies from conversation to conversation, leading to a context of variable size. To

eliminate this complication, when constructing or accessing the model describing the kth row of chronogram $\mathbf{Q}$, the remaining $K-1$ rows (representing the kth participant's interlocutors) are collapsed via an inclusive-OR operation, to provide a single "all interlocutors" row. This results in a conditioning history of τ frames of the kth participant, and τ frames of context describing whether any of the kth participant's interlocutors were speaking at instant $t-\tau$ (Laskowski et al., 2011b).

The above method yields a history duration which is independent of K, and lends itself easily to N-gram modeling. The elements of the conditioning history are marshalled into a one-dimensional order, and counts are accumulated as elsewhere for N-grams. This results in a maximum-likelihood (ML) model $p_A\left(q|h\right)$ for a sequence denoted A, with $q \in \{\square, \blacksquare\}$ and h the conditioning history. In (Laskowski et al., 2011b), such models were interpolated with lower-order (smaller-τ) models (Jelinek and Mercer, 1980), yielding smoothed models $\tilde{p}_A\left(q|h\right)$. In the absence of smoothing, as in the current work, the order of the elements of the $(2\times\tau)$-length history is unimportant, provided it is fixed.

3.3 Supervised Modeling

In supervised modeling, a model A is constructed from one or more conversation sides attributed to the same speaker, and then that model is applied to a conversation side B whose speaker is unknown. In this case, a commonly used score between generative model A and sequence B is the *average negative log-likelihood* of the sequence given the model, which is also known as the *conditional cross entropy*:

$$\begin{aligned} & H\left(p_B\left(q|h\right)|\tilde{p}_A\left(q|h\right)\right) \\ & \quad = \quad -\sum_{h,q} p_B\left(h,q\right)\log\tilde{p}_A\left(q|h\right) \ , \end{aligned} \tag{1}$$

where $p_B\left(h,q\right)$ are the ML joint probabilities observed in sequence B. Equation 1 is often normalized by subtracting the *conditional entropy* (Cover and Thomas, 1991),

$$\begin{aligned} & H\left(p_B\left(q|h\right)\right) \\ & \quad = \quad -\sum_{h,q} p_B\left(h,q\right)\log p_B\left(q|h\right) \ . \end{aligned} \tag{2}$$

yielding the *conditional relative entropy* or *conditional Kullback-Leibler divergence* (Cover and Thomas, 1991):

$$\begin{aligned} & D_{KL}\left(p_B\left(q|h\right)\|\tilde{p}_A\left(q|h\right)\right) \\ & \quad = \quad \sum_{h,q} p_B\left(h,q\right)\log\frac{p_B\left(q|h\right)}{\tilde{p}_A\left(q|h\right)} \ . \end{aligned} \tag{3}$$

For example, in the context of stochastic turn-taking models, Equation 1 was successfully used with zero-normalization of scores (Laskowski, 2014).

3.4 Unsupervised Modeling

In the unsupervised case, a score does not normally compare a sequence B to a model A, but rather a sequence A to a sequence B (or, alternately, a model trained on sequence A to a model trained on sequence B). Because of this symmetry, it is desirable for the score itself to be symmetric; the conditional Kullback-Leibler divergence in Equation 3 does not exhibit this quality and, additionally, is unbounded. It is therefore customary to compute the conditional Jensen-Shannon divergence (Lin, 1991), which for two equal-weight conditional probability models p_A and p_B is given by

$$\begin{aligned} & D_{JS}\left(p_A\left(q|h\right)\|p_B\left(q|h\right)\right) \\ & \quad \equiv \quad \frac{1}{2}\,D_{KL}\left(p_B\left(q|h\right)\|p\left(q|h\right)\right) \\ & \quad + \quad \frac{1}{2}\,D_{KL}\left(p_A\left(q|h\right)\|p\left(q|h\right)\right) \ . \end{aligned} \tag{4}$$

Here, $p\left(q|h\right)$ is the "joint-source" (ie. A and B) model; (El-Yaniv et al., 1997) showed that for models of conditional probability, its form is

$$\begin{aligned} p\left(q|h\right) \quad &= \quad \lambda_A\left(h\right)\cdot p_A\left(q|h\right) \\ &+ \quad \lambda_B\left(h\right)\cdot p_B\left(q|h\right) \ , \end{aligned} \tag{5}$$

namely that it is the linear interpolation of the two single-source models, with weights given by their relative probabilities of the occurrence of the context h:

$$\lambda_A\left(h\right) \quad = \quad \frac{p_A\left(h\right)}{p_A\left(h\right)+p_B\left(h\right)} \tag{6}$$

$$\lambda_B\left(h\right) \quad = \quad \frac{p_B\left(h\right)}{p_A\left(h\right)+p_B\left(h\right)} \ . \tag{7}$$

The *Jensen-Shannon distance*, a score which is both bounded and symmetric, is given by

$$d_{A,B} \quad \equiv \quad \sqrt{D_{JS}\left(p_A\left(q|h\right)\|p_B\left(q|h\right)\right)} \ . \tag{8}$$

Table 1: Leave-one-out (LOO) modified-KNN classification accuracies, using Jensen-Shannon distances between STT models of individual conversation sides in DATASET. K specifies the maximal number of neighbors; τ is the number of 100-ms frames of conditioning history. Each frame contains 2 bits of information: whether the modeled-side participant was speaking, and whether any of that participant's interlocutors were speaking.

K	τ							
	1	2	3	4	5	6	7	8
1	0.37	0.44	0.56	0.54	0.47	0.37	0.18	0.09
3	0.36	0.53	0.51	0.55	0.48	0.37	0.16	0.09
5	0.40	0.53	0.59	0.58	0.49	0.34	0.15	0.07
7	0.40	0.54	0.59	0.57	0.49	0.33	0.16	0.07
9	0.41	0.54	0.57	0.57	0.50	0.33	0.13	0.07
11	0.43	0.55	0.59	0.57	0.52	0.33	0.15	0.08
13	0.43	0.54	***0.60***	0.57	0.54	0.34	0.15	0.09
15	0.45	0.54	***0.60***	0.58	0.54	0.35	0.18	0.10
17	0.45	0.54	0.59	0.59	0.55	0.36	0.20	0.13
19	0.45	0.55	***0.60***	0.58	0.54	0.38	0.21	0.13
25	0.44	0.53	0.57	0.57	0.53	0.38	0.21	0.13

3.5 Modified Nearest-Neighbor Classification

A central goal of the current work is the determination of whether two sequences, produced by the same person in different conversations, are more proximate than are two sequences produced by two different persons. One answer to this question can be provided by classifying sequences based on their proximity, of which the formalization is known as K-nearest neighbor classification (Fix and Hodges, 1951). The input to the algorithm is a symmetric, zero-diagonal distance matrix D, whose entries are pair-wise distances.

Here, a modified version of the algorithm is employed. If the speaker g of the side being classified is known to have produced only $N_g - 1$ other sides in the collection of sides under study, then K is limited to $N_g - 1$ for that classification trial. The use of such side information may be perceived as unfair; however, the aim is diagnostic, and no effort has been made in the current work to normalize the distances in D for local density differences. In addition, it makes little sense to penalize an analysis for those trials whose speakers produced no other sides in DATASET (cf. Section 2). The results of such a diagnostic test can be usefully compared to the outcome of random guessing under the same circumstances.

An alternative approach, consisting of applying clustering to the distance matrix, was also tried; the results yielded similar (albeit more difficult to disentangle) results and are not presented due to space constraints.

3.6 Multidimensional Scaling

Finally, multidimensional scaling (MDS; cf. (Borg and Groenen, 2005) for example) was applied in an attempt to embed models in a low-dimensional space and to facilitate visual analysis. The experiments used the `smacofSym()` function (de Leeuw and Mair, 2009) implementation in R.

4 Results

For a given $\tau \in [1, 2, 3, \ldots, 8]$, each conversation side q_n of the $N = 197$ sides in DATASET was used to train a side-specific maximum likelihood (ML) model θ_n. The distance between every pair of models was then computed using Equation 8, leading to a symmetric, zero-diagonal distance matrix $D \in \mathbb{R}_+^{197\times197}$.

4.1 Diagnostic Classification

D was then used within the modified K-nearest neighbor participant-identity classification framework described in Section 3.5. The achieved accuracies are shown in Table 1.

As can be seen, the highest accuracies are obtained for $\tau \in [2, 3, 4, 5]$ with $K > 7$, with an absolute maximum from among those explored of 60%, at $\tau = 3$ and $K = 15$. This is considerably in excess of 11%, the accuracy

Table 2: LOO modified-KNN classification accuracies, using distances computed following multidimensional scaling (MDS) of the distances between STT models of individual conversation sides in DATASET, to 5 dimensions. Compare to Table 1.

K	τ							
	1	2	3	4	5	6	7	8
1	0.37	0.47	0.49	0.56	0.59	0.54	0.55	0.47
3	0.39	0.49	0.57	0.58	0.62	0.61	0.58	0.48
5	0.39	0.46	0.61	0.62	0.65	0.62	0.60	0.52
7	0.40	0.48	0.59	0.63	0.66	0.61	0.59	0.53
9	0.43	0.51	0.58	0.63	0.66	0.62	0.56	0.54
11	0.43	0.49	0.58	0.62	0.68	0.61	0.59	0.53
13	0.43	0.49	0.58	0.63	0.68	0.60	0.60	0.52
15	0.45	0.51	0.57	0.64	0.69	0.61	0.59	0.52
17	0.44	0.52	0.60	0.66	***0.70***	0.63	0.59	0.54
19	0.45	0.53	0.60	0.65	0.69	0.62	0.58	0.54
25	0.44	0.53	0.59	0.65	0.68	0.63	0.59	0.53

achieved by random guessing with the DATASET priors. This result corroborates the findings in (Laskowski, 2014), that participant identities can frequently be inferred from STT models; the difference with (Laskowski, 2014) is that in the latter work, models were trained on same-person *sets* of sides in a training portion of the data, rather than on individual sides, and that the asymmetric conditional cross entropy (Equation 2, with zero-normalization) was used rather than Jensen-Shannon divergence (Equation 4).

4.2 Diagnostic Classification after Scaling

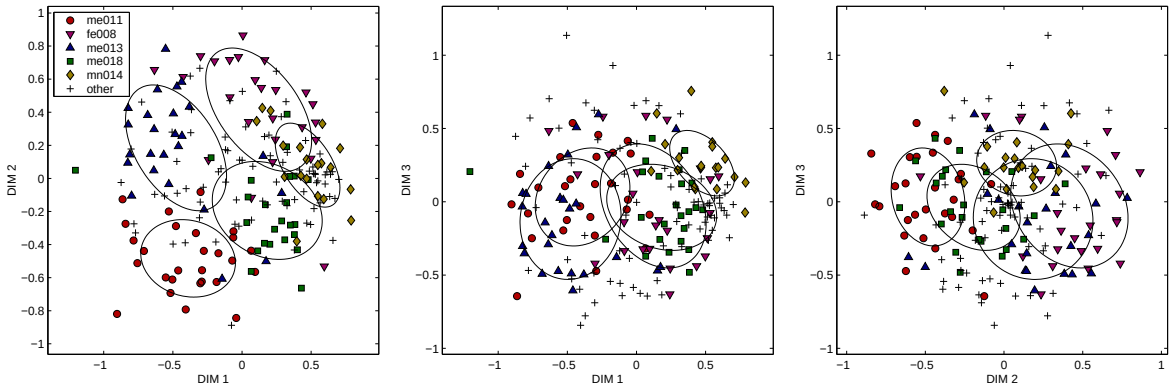
The computed pair-wise Jensen-Shannon distances lie in a space of unknown effective dimensionality; the determination of that effective dimensionality is one of the implicit aims of the current work. To this end, the distances were embedded in a fixed-dimensionality subspace, using multidimensional scaling (MDS) as described in Section 3.6. All 19306 pair-wise distances comprising D were then re-computed from the MDS-derived positions, and the diagnostic experiment of Section 4.1 was repeated. The results for a 5-dimensional subspace are shown in Table 2.

As can be seen, relative to Table 1, MDS to 5 dimensions actually increases the attainable classification accuracy, to 70% at $\tau = 5$ and $K = 17$. This suggests that there is considerable noise in the distance estimates, and that scaling effectively collapses some of that variability. The accuracy-maximizing number of dimensions, whose identification is beyond the scope of the current work, is expected to be specific to any particular data set. However, it is notable that for DATASET this "elimination of unwanted variance" occurs for the higher-complexity ($\tau > 2$) models; distances computed using these are more likely to be noisy that those computed using simpler models, for fixed conversation-side durations. Since the $\tau = 8$ context contains the $\tau = 5$ context, this suggests that the duration of the conversations studied here, between 22.8 and 74.5 minutes, may be insufficient to infer robust long-conditioning-history models.

Similar experiments were performed after MDS scaling to each of $\{4, 3, 2, 1\}$ dimensions. The results are not shown due to space constraints. A summary of the maximum achieved accuracy in each case is depicted in Figure 2.

The figure shows that with each reduction of dimensionality of the embedding subspace, by one additional dimension, the maximum achievable accuracy falls by an increasing amount. Although for a one-dimensional subspace the accuracy of 35% is still considerably above chance (11%), it is already (just) less than halfway to the accuracy achieved without scaling (60%).

At 3 dimensions, the accuracy of 58% is almost the same as that achieved without scaling; it occurs at $\tau = 6$ and $K = 17$ (not shown). This suggests that the relative magnitudes of the distances are preserved in a continuous small-dimensional space, and may have implications for understanding what STT models actually learn. For example, each of the dimensions may be strongly correlated



(a) dimensions 1 & 2 (b) dimensions 1 & 3 (c) dimensions 2 & 3

Figure 3: Positions of 197 models, each of one conversation side in DATASET, as inferred using a Jensen-Shannon distance matrix and multidimensional scaling (MDS) to 3 dimensions. Sides produced by the five most frequently-occurring persons (cf. Section 2) are identified explicitly, together with ellipses representing projections of the corresponding 50% error ellipsoid.

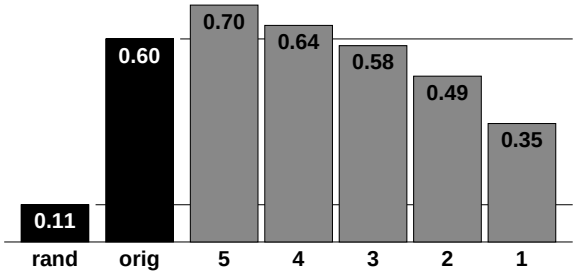


Figure 2: Maximum achieved LOO modified-KNN classification accuracies, using distances computed following MDS down to $[5, 4, 3, 2, 1]$ dimensions of the distances between STT models of individual conversation sides in DATASET. The accuracies are compared to the maximum accuracy achieved using unscaled distances ("orig") and random guessing with actual LOO priors ("rand").

with an independently measurable human trait or role trait. In that case, such traits could be used to index STT models, for both generation and recognition purposes in multi-party conversational settings.

4.3 Model Subspace Visualization

It is serendipitous that, for the data set under investigation, three dimensions suffice to yield a good approximation of the accuracy achievable without scaling. A three-dimensional space is considerably easier to inspect visually, and to understand, than are higher-dimensional spaces. Figure 3 shows the MDS-derived locations, two dimensions at a time. The 197 datapoints, representing models of individual conversation sides, are seen to comprise a cloud with heterogenous, locally clumpy density. The determinant of the total scatter matrix, given these inferred positions, is 2.74×10^3.

The determinants of the between-class scatter matrix and the within-class scatter matrix, given the model positions shown in Figure 3, are 3.29×10^3 and 2.86×10^3, respectively. It appears from these numbers that the variability between different-person sides is on average larger than the variability between same-person sides, which in turn suggests that people exhibit low variability — even across longitudinal spans of many months — relative to what differentiates them from others.

5 Discussion

5.1 Intra-Person Variability

It is relevant to try to determine whether the variability observed among models of the same person are due to actual variability of behavior or to measurement error. One source of measurement error could be the relative duration of conversations, leading to unequally (under)trained models. Figure 4 depicts the five most frequent participants in DATASET, at the same positions as in Figure 3(a), with marker size indicative of the duration of observation.

It can be seen that, broadly, shorter-duration conversations yield models which lie at the periphery of the error ellipses. This indicates that — were conversations longer or models more par-

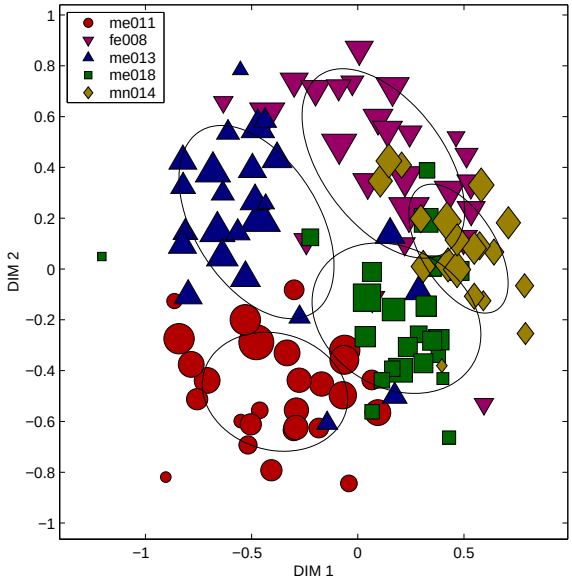


Figure 4: Replication of Figure 3(a) with marker size linearly proportional to the duration of conversation from which each side is drawn. Sides for only the top five most frequent participants shown.

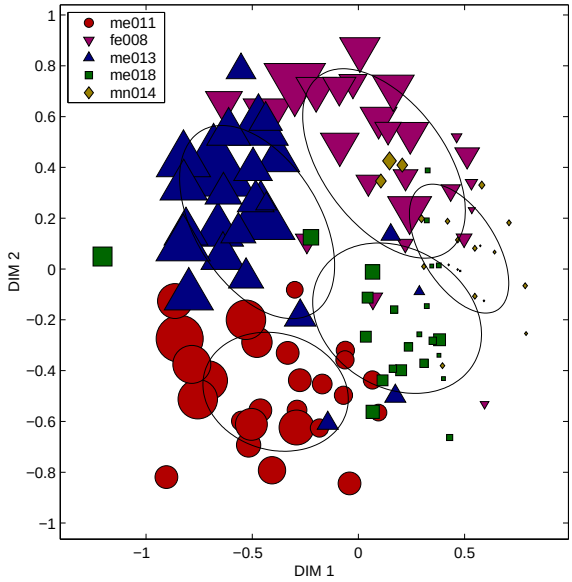


Figure 5: Replication of Figure 3(a) with marker size linearly proportional to the amount of speech observed for each side. Sides for only the top five most frequent participants shown.

simonious — the resulting error ellipses (shown unchanged from Figure 3(a) in Figure 4) may be tighter, and thereby even more discriminative.

A second potential source of intra-person variability may be not just the duration of observation (i.e. the duration of conversation), but how talkative a person is during a specific conversation. Although the models employed here make no mathematical distinction between speaking and not speaking, in multi-party turn-taking the average participant speaks for only a minority of time, making speaking (versus not speaking) a distinctively marked behavior. Figure 5 is like Figure 4, but marker size is indicative of the amount of speech observed for each side.

Figure 5 shows that points lying in the bottom right of the figure represent low quantities of speech per side, globally. This appears to be true for individual speakers separately, particularly for the top three most frequent participants (and `me013` most markedly). Since the ellipses appear cigar-shaped, fanning out from the bottom right, these observations suggest that when given the opportunity to speak a lot, participant models "move" to the upper left where they may be even further apart. They also suggest that a quantity encoded in the plane of the first and second MDS dimensions ("DIM1" and "DIM2" in the figure) is the proportion of speech produced by each person, or their "talkativity".

5.2 Inter-Person Variability

A source of established (Laskowski et al., 2008) variability in turn-taking models trained using the ICSI Meeting Corpus is the relative seniority of participants within a group. (Laskowski et al., 2008) used the self-reported `Education` level. Figure 6 retains the topology shown in Figure 3(a), but markers represent the educational level of individual participants in DATASET. It can be seen that students (`Undergrad` and `Grad`) occupy exclusively the lower half in the diagram, while `Postdoc` and `Professor` are found predominantly in the upper half, but in separate clusters. Persons of type `PhD` exhibit no such leanings.

Figure 6 suggests that education level is indeed discriminated by the STT-model topology inferred via MDS. (Laskowski et al., 2008) observed that despite the fact that persons of type `Professor` spoke a lot, they appeared to avoid overlap with persons of type `Undergrad`. Such tendencies are most likely the result of social roles within the organization, and not of educational level per se, but role and education level are probably very correlated in an academic setting. It may be tentatively concluded that the ("DIM 1","DIM 2") plane also encodes, in addition to each person's "talkativity" (cf. Subsection 5.1), their tendency to initiate and

terminate talk in overlap.

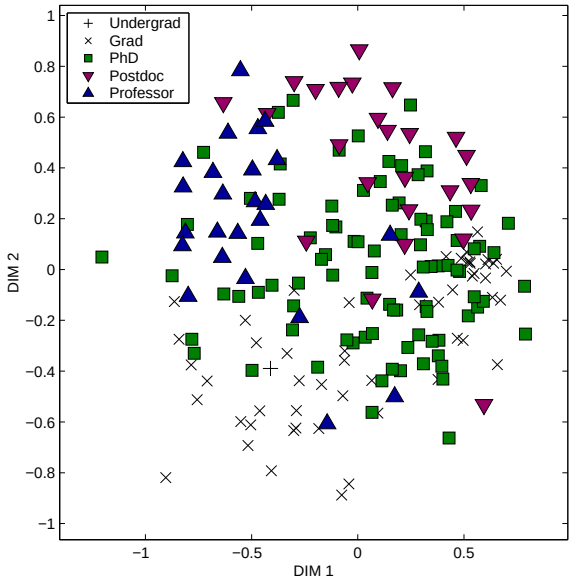


Figure 6: Replication of Figure 3(a) with marker shape denoting the self-reported education level of each side.

It should be noted that, unlike the measurement of intra-person variability, the measurement of inter-person variability is likely a function of the size of the group of people studied. As described in Section 2, the group considered here consists of 15 individuals, some of which participated in only a handful of conversations. For larger groups, it can be expected that — if models represent interaction styles — inter-person variability under a fixed model order and a fixed observation duration will decrease, since nothing a priori prevents multiple individuals from interacting using the same or similar-enough style. Since intra-person variability is independent of the number of other persons considered, it is expected to remain constant under group resizing. The ratio of the inter-person variability to the intra-person variability is therefore likely to decrease with increasingly larger group sizes, when the model complexity and observational duration remain constant.

5.3 Training Speaker-Independent Models

That within-person SST-model variability can be smaller than between-person variability, as discovered in the dataset used in the current study, has important consequences for training broad STT models, intended to be applicable to a wide variety of domains and conversational interaction styles. The results presented indicate that including more training data, without careful consideration of its interaction-style content, may bias the model towards the styles present in the training data and therefore away from the styles in test data — since they *can* be so different. In this sense, the results corroborate earlier, similar findings for domain and topic variability in language modeling within automatic speech recognition.

5.4 Potential Impact and Applications

Over and above the immediate recommendations for the training of STT models, the results obtained in the current study may have several consequences for at least three research areas.

An understanding of the contexts in which participants to conversation choose to vocalize can usefully inform the construction of speaker diarization systems. Current state-of-the-art diarization technology, as used in the transcription of far-field recordings of multi-party meetings, oversegments the temporal support of the recorded track and then performs agglomerative hierarchical clustering using spectral or voice-print similarity. The prior knowledge used in these systems consists of minimal duration constraints on intervals of single-party talk, as well as the assumption that each instant is associated with exactly one participant speaking. The detection of overlap (or of simultaneous vocalization by more than one speaker), where performed, is generally treated as a post-processing step. Information regarding consistent, participant-specific tendencies in the temporal deployment of talk — the subject of the current study — do not currently feature in any way in the assumptions or priors of today's diarization systems.

Second, dialogue system design can benefit from the results presented, particularly those systems which are conversational and whose behavior is intended to be more natural than that of simple human-query-driven information portals. The confirmation that humans exhibit self-consistency in their temporal deployment of speech, which also makes them different from other people, means that the detection of their style and an orientation to it will result in better predictions, requiring fewer resolutions. If that orientation is perceivable to the human user, the system may appear to the user as more human itself. An additional dimension of human-likeness may be inadvertently communicated by the system if it has its own, self-

consistent and differentiable style, which is syntonic with its designed conversational role.

Finally, the results in this study have bearing on the design of diagnostic tools for social psychology, the domain for which STT models were originally invented (Chapple, 1949; Jaffe et al., 1967). (Chapple, 1949) was concerned with the measurement of conversational traits correlated with work performance, whereas (Jaffe et al., 1967) treated clinical settings. A considerable amount of research in this area had been conducted in the 1970s and 1980s, primarily in the detection of traits or conditions. However, the models were first-order Markovian (corresponding to $\tau = 1$ in the current work) and often relying on analysis frames as small as 20 ms. The findings presented here indicate that useful speaker-discriminating information is contained as far back as 500 ms (with frames of 100 ms and $\tau = 5$, cf. Subsection 4.2), even when models are trained on single conversations which are as short as 22 minutes long. The obtained results may warrant a re-opening of earlier investigations into diagnostic tools for the health industry.

6 Conclusions

That people exhibit a degree of consistency in their conversational behavior agrees with common sense, and should not be particularly surprising. A number of earlier works have successfully correlated identity with turn-taking preferences (Jurafsky et al., 2009; Grothendieck et al., 2011). What the analyses in the current work show — and which is surprising — is that this consistency is present even in the very shallow representation implicit in the so-called stochastic turn-taking models. In this representation, words, boundaries, durations, and prosody are markedly absent; only the frame-level occurrence of party-attributed speech activity is captured, and a definition of "turn" is neither needed nor used. Specifically, results indicate that, for conversations whose duration is 40-minutes on average, longitudinally speaker-discriminative models can be learned for a conditioning history which is only 10 bits long: whether the modeled speaker, and *any* of their interlocutors, were speaking in each of the 5 most recent 100-ms frames. The current study has shown that under these conditions, for groups of 15 people like the ICSI `Bmr` group, the inferred models exhibit greater between-person variability than within-person variability. The conversants under study appear to have behaved self-consistently, across disparate longitudinal observations, in terms of their turn-taking preferences.

The current experiments also demonstrated that a conversation-side embedding in *three* dimensions approximately recovers the Jensen-Shannon distances between 10-bit-context STT models. In this embedding, between-person variability was shown to be smaller for longer conversations, implying that over time people can be observed to converge on interaction styles which are even more self-consistent. Although it is premature to unambiguously ascribe meaning to each of the three dimensions obtained using the ICSI `Bmr` data, jointly they appear to encode: (1) the proportion of conversation-time spent talking; (2) the inclination to initiate and terminate overlap with others; and (3) role-specific behaviors exhibited by members of a hierarchy (with — in the current work — positions within that hierarchy closely correlated with self-reported education level).

The presented work suggests the possibility of inference of speaker-characterizing conversational interaction styles, as well as the indexing of such interaction styles by points in an embedding space consisting of only a few continuous dimensions. It has immediate bearing on the training of intentionally broad, speaker-independent STT models. Finally, the work has the potential to usefully impact the design of speaker diarization algorithms for multi-human conversation settings, of human-like conversational dialogue systems, and of diagnostic software for the health industry.

7 Acknowledgments

This work was funded in part by the Riksbankens Jubileumsfond (RJ) project *Samtalets Prosodi*. Computing resources at Carnegie Mellon University were made accessible courtesy of Qin Jin and Florian Metze.

References

I. Borg and P. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*. Springer.

P. Brady. 1969. A model for generating on-off speech patterns in two-way conversation. *The Bell System Technical Manual*, 48(9):2445–2472.

E. Chapple. 1949. The Interaction Chronograph: Its evolution and present application. *Personnel*, 25(4):295–307.

T. Cover and J. Thomas, 1991. *Elements of Information Theory*, chapter Entropy, Relative Entropy and Mutual Information (Chapter 2), pages 12–49. John Wiley & Sons, Inc.

J. de Leeuw and P. Mair. 2009. Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, 31(3):1–30.

R. El-Yaniv, S. Fine, and N. Tishby. 1997. Agnostic classification of Markovian sequences. In *Proc. Advances in Neural Information Processing Systems (NIPS) 10*, pages 465–471, Denver CO, USA.

E. Fix and J. Hodges. 1951. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical report, USAF School of Aviation Medicine, Randolph Field TX, USA.

J. Grothendieck, A. Gorin, and N. Borges. 2011. Social correlates of turn-taking style. *Comput. Speech Lang.*, 25(4):789–801, October.

Jr. J. Dabbs and R. Ruback. 1987. Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology*, 20:123–169.

J. Jaffe, S. Feldstein, and L. Cassotta. 1967. Markovian models of dialogic time patterns. *Nature*, 216:93–94.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. In *Proc. 28th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 364–367, Hong Kong, China.

F. Jelinek and R. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands.

D. Jurafsky, R. Ranganath, and D. McFarland. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638–646, Boulder CO, USA.

K. Laskowski, M. Ostendorf, and T. Schultz. 2008. Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In *Proc. 9th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, Columbus OH, USA.

K. Laskowski, J. Edlund, and M. Heldner. 2011a. Incremental learning and forgetting in stochastic turn-taking models. In *Proc. 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2065–2068, Firenze, Italy.

K. Laskowski, J. Edlund, and M. Heldner. 2011b. 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)*, pages 5600–5603, Praha, Czech Republic.

K. Laskowski. 2010. Modeling norms of turn-taking in multi-party conversation. In *Proc. 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 999–1008, Uppsala, Sweden.

K. Laskowski. 2012. Exploiting loudness dynamics in stochastic models of turn-taking. In *Proc. 4th IEEE Workshop on Spoken Language Technology (SLT)*, pages 79–84, Miami FL, USA.

K. Laskowski. 2014. On the conversant-specificity of stochastic turn-taking models. In *Proc. 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2026–2030, Singapore.

J. Lin. 1991. Divergence measrures based on the Shannon entropy. *IEEE Trans. Information Theory*, 37(1):145–151.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proc. 5th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Cambridge MA, USA.

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