AN INFORMATION-THEORETIC FRAMEWORK FOR AUTOMATED DISCOVERY OF PROSODIC CUES TO CONVERSATIONAL STRUCTURE

K. Laskowski

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

A. Hjalmarsson

KTH Speech, Music and Hearing
Stockholm, Sweden

ABSTRACT

Interaction timing in conversation exhibits myriad variabilities, yet it is patently not random. However, identifying consistencies is a manually labor-intensive effort, and findings have been limited. We propose a conditonal mutual information measure of the influence of prosodic features, which can be computed for any conversation at any instant, with only a speech/non-speech segmentation as its requirement. We evaluate the methodology on two segmental features: energy and speaking rate. Results indicate that energy, the less controversial of the two, is in fact better on average at predicting conversational structure. We also explore the temporal evolution of model "surprise", which permits identifying instants where each feature's influence is operative. The method corroborates earlier findings, and appears capable of large-scale data-driven discovery in future research.

Index Terms— interaction modeling, speaking rate, conditional mutual information, neural networks, automated discovery.

1. INTRODUCTION

Interaction timing in conversation exhibits myriad variabilities, yet it is patently not random. Participants seem to know just when to begin speaking, appear secure in their ability to continue uncontested when they pause, and seem to indicate when others may speak without explicit mention [1]. However, identifying the consistencies that humans must be exploiting has proven difficult. This appears to be due to problems with representation, annotation effort, and complexity of modeling; often the three problems are intertwined. As a result, limited findings are available for a limited number of theory-driven event types (e.g. "back-channels", or "turn" terminations), in limited domains, languages, and corpora.

To address this sparseness of findings, we propose a method for the automated, data-driven discovery of how frame-level signal-derived features may contribute to interactional timing at *any* instant of *any* dialogue. The method is theory-agnostic and enables easy comparison of features in arbitrarily large collections of data. We ask two specific questions:

Q1: Can the method assess the degree of influence that a feature has on conversational structure?

Q2: Can the method point to conversational contexts in which the feature's influence is operative?

To test the method, we deliberately explore the influence of two features — one whose effect is well-accepted in the literature and a second whose impact is more controversial. The well-accepted feature is energy, a popular correlate of speaking loudness or intensity. Many studies have reported that there is a fall in intensity at the ends of "turns" (e.g. [2, 3, 4]), suggesting that negative-slope energy contours could be exploited by other participants when deciding when to begin talking.

The more controversial feature we explore is speaking rate. It has been proposed that "a draw on the final syllable or on the stressed syllable of a terminal clause" has a turn-yielding effect [5], suggestiong that ends of talkspurts exhibit reduced speech rates. However, subsequent studies have found no clear evidence of such a relationship. [6] has claimed that increased phrase-final lengthening has turn-yielding functions in Tyneside English. [7] and [3], on the other hand, present results that suggest that lengthening occurs in all phrase-final positions, but that segmental lengthening prior to speaker change is significantly shorter than before pauses in turn-medial position. A more recent study suggests that utterance-final lengthening has a turn-holding effect on Swedish listeners, but no such effect was found for English listeners [8]. In summary, despite some agreement on the influence of intensity or energy, no broadly observable and uncontestable findings appear to be had for the influence of speaking rate. We note also that the above cited works rely on annotations of syllables, stressed syllables, terminal clauses, "turns", phrases, and/or turn-medial positions, and that these may be hard to obtain automatically, limiting the scope of investigation.

The present article comprises two contributions. First, we apply a previously proposed stochastic turn-taking (STT) model [9] to a correlate of speaking rate known as Mel-spectral flux (MSF) [10]; [9] had explored energy as a correlate of speaking loudness. The STT experiments demonstrate that our method does in fact permit comparison, answering *Q1* in the affirmative. In particular, it demonstrates that MSF is helpful in predicting incipient speech, but not as helpful as energy. Second, we present a previously unexplored use of STT models: the analysis of the trajectory of model "surprise" as the conversation proceeds. Such analysis — answering *Q2* in the affirmative — shows that models are often "surprised" at speakers' speech/non-speech boundaries and at listeners' non-speech during their interlocutors' pauses. The proposed method thereby appears capable of unmediated automatic discovery of which features are most operative at which instants in mitigating the surprise of the incipient conversational future.

2. FRAMEWORK

The structure of a conversation, in terms of the timing of participants' decisions to deploy non-speech or speech, is most easily represented as a chronogram $\mathbf{Q} \in \{\square, \blacksquare\}^{K \times T}$ [11]. "$\square$" and "$\blacksquare$" represent non-speech and speech, respectively; K is the number of participants, and T is the number of frames which results from a discretization of time. We perform this discretization with non-overlapping frames of 100-ms, ensuring generally sub-syllable granularity. In the current work, K is identically 2. An example of a

chronogram is

$$\mathbf{Q} \quad = \quad \left[\ldots \begin{array}{c} \blacksquare\blacksquare\blacksquare\blacksquare\square\square\square\square\blacksquare \\ \square\square\square\blacksquare\blacksquare\blacksquare\blacksquare\square\square \end{array} \ldots \right] \quad . \tag{1}$$

Our framework considers the probability distribution of chronograms. For any $\mathbf{Q}$,

$$P(\mathbf{Q}) \quad \doteq \quad \prod_{t=t}^{T} P\left(\mathbf{q}_t \middle| \mathbf{q}_{t-1}^{t-\tau}\right) \quad . \tag{2}$$

Namely, we factor the probability in time, in the same way that the probability of a reference word sequence is decomposed in language modeling (e.g., [12], from which subscript notation in Equation 2 is borrowed); here τ is the number of frames retained in the conditioning context. Furthermore, we consider the K participants *conditionally independent* at any instant t, given their joint past $\mathbf{q}_{t-1}^{t-\tau}$:

$$P\left(\mathbf{q}_t \middle| \mathbf{q}_{t-1}^{t-\tau}\right) \quad \doteq \quad \prod_{k=1}^{K} P\left(\mathbf{q}_t\left[k\right] \middle| \mathbf{q}_{t-1}^{t-\tau}\right) \quad . \tag{3}$$

We estimate the right-hand-side factors, i.e.

$$\mathbf{y}_t\left[k\right] \quad = \quad P\left(\mathbf{q}_t\left[k\right] \middle| \mathbf{q}_{t-1}^{t-\tau}\right) \quad , \tag{4}$$

by training a feed-forward neural network $\mathcal{M}$ using a TRAINSET of conversations. The network contains one hidden layer with $J \in \left[1, 2, 4, 8, 16, 32, 64\right]$ units; the dot-product followed by a sigmoid provides the activations for all units. The appropriate error for a neural network whose single output is limited to the unit interval is the cross-entropy error [13], which we minimize using scaled conjuge gradient search. The sum of that error over all the $\mathbf{Q}$'s in TRAINSET is the cross-entropy $H\left(\left\{\mathbf{Q}\right\} \middle| \mathcal{M}\right)$ [14].

The extension to some additional feature $\mathbf{f}$ — provided it can be massaged into a matrix $\mathbf{F}$ of size $K \times T$, just as $\mathbf{q}$ was — is straightforward. Instead of Equation 4, the factors we estimate are

$$\mathbf{y}_t\left[k\right] \quad = \quad P\left(\mathbf{q}_t\left[k\right] \middle| \mathbf{q}_{t-1}^{t-\tau}, \mathbf{f}_{t-1}^{t-\tau}\right) \quad . \tag{5}$$

This leads to a distribution $P\left(\mathbf{Q} \middle| \mathbf{F}\right)$, which is still a "prior" distribution over $\mathbf{Q}$ provided that we do not peek at future instants in $\mathbf{F}$ when estimating earlier speech activity states in $\mathbf{Q}$.

For any $\mathbf{Q}$ and its associated $\mathbf{F}$, the difference between the outputs of the $\mathbf{F}$-conditioned and the not-$\mathbf{F}$-conditioned networks is the conditional mutual information [14]

$$I\left(\mathbf{Q}, \mathbf{F} \middle| \mathcal{M}\right) \quad = \quad H\left(\mathbf{Q} \middle| \mathcal{M}\right) - H\left(\mathbf{Q} \middle| \mathbf{F}, \mathcal{M}\right) \quad , \tag{6}$$

which directly measures the impact of $\mathbf{F}$ on *future* phenomena in $\mathbf{Q}$, given model $\mathcal{M}$. In what follows, we will refer to $H\left(\cdot\right)$ as the cross entropy rate or "surprise", expressed in bits per 100-ms.

3. EXPERIMENTS

3.1. Data

Experiments are conducted using the Switchboard-1 Corpus, as rereleased in 1997 [15]. It consists of 2435 telephone conversations, each approximately 10 minutes in duration. The corpus was divided into three speaker-disjoint sets in [16], such that TRAINSET, DEVSET, and TESTSET consist of 762, 227, and 199 conversations, respectively. During that process, it was not possible to allocate 1247 conversations because their two speakers had already been placed in different sets. The available forced alignments [17] for both conversation sides were used to construct $\mathbf{Q}$; in particular, a $\mathbf{Q}$ frame at instant t for participant k was declared as ■ if k was speaking for at least 50 ms of the tth 100-ms interval, and □ otherwise.

3.2. Prosodic Features

The current article compares two types of features, energy and speaking rate. The matrix $\mathbf{F}$ (from Equation 6) for energy will be denoted $\mathbf{E}$; its entries were computed by pre-emphasizing the audio, squaring the amplitudes, applying a rectangular 100-ms window, and taking the logarithm of the result.[1]

As a frame-level correlate of speaking rate, we computed Mel-spectral flux (MSF) [10]. This was done by considering two windows, as shown in Figure 1. The cosine distance is computed between the Mel-spectra of the two windows after signal pre-emphasis, and the negative logit function is applied to the result. The pair of windows, whose joint duration is $2t_{ext} + t_{sep}$, is then shifted to the right by 8 ms; an average of n such consecutive measurements is taken to be the value of each cell of $\mathbf{R}$, our second variant of $\mathbf{F}$. n is given as $\lfloor t_{ave} / \left(2t_{ext} + t_{sep}\right) \rfloor$, where t_{ave} is an integration interval whose right edge coincides with the right edge of the 100-ms frame in $\mathbf{R}$. Interested readers are directed to [10] for the motivations of this design.

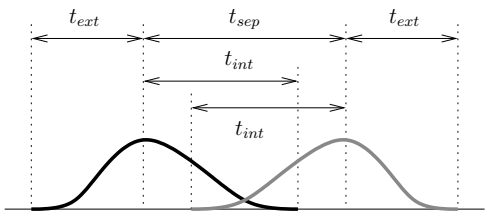

Fig. 1. An MSF subframe consisting of a left and a right window, in black and gray, respectively.

In adapting the MSF definition in [10] to the current work, we optimized t_{int}, t_{sep}, t_{ext}, and t_{ave} by minimizing the TRAINSET cross entropy. Their values are 30 ms, 84 ms, 18 ms, and 350 ms, respectively.

3.3. Prediction Performance using Individual Feature Types

Prediction of incipient speech activity was performed individually using $\mathbf{Q}$, $\mathbf{E}$, and $\mathbf{R}$. The cross entropies achieved are shown for all three datasets TRAINSET, DEVSET, and TESTSET in Table 1. As can be seen, the most predictive feature (at our frame rate of 10 Hz) is $\mathbf{Q}$. When used by themselves, $\mathbf{E}$ appears considerably better than $\mathbf{R}$ for predicting future $\mathbf{Q}$. We observe no bias towards lower cross entropies on TRAINSET, suggesting an absence of overfitting.

To exclude the possibility that $\mathbf{R}$ is insufficiently correlated with speaking rate, we assessed two suprasegmental speaking rate features on the same task. The first is a syllable rate $\mathbf{S}^{(0.1)}$, which counts the number of syllable nuclei in the current frame's 100-ms support. The nuclei were detected automatically from the audio, based on intensity peaks in dB during voiced segments, using a method introduced by [18]. We also extended the integration interval to $\left\{200, 300, 400\right\}$ ms back from the right edge of the current frame, leading to variants $\mathbf{S}^{(0.2)}$, $\mathbf{S}^{(0.3)}$, and $\mathbf{S}^{(0.4)}$, respectively. The second suprasegmental rate feature is word rate $\mathbf{W}^{(0.1)}$, obtained by

[1] In this way it differs from the energy feature computed in [9]. There, a 200-ms Hamming window was used, the right-side taper of which peeked 50 ms into the future 100-ms frame whose speech state was being predicted. To our surprise, the current correction actually improves results, by negligibly reducing the cross entropies in all cases.

Table 1. Cross entropies for three feature types. J^* is the number of hidden units, selected by minimizing DEVSET cross entropy.

Feature	J^*	TRAINSET	DEVSET	TESTSET
$\mathbf{Q}$	32	0.2788	0.274321	0.275023
$\mathbf{E}$	64	0.363512	0.353707	0.35872
$\mathbf{R}$	64	0.547963	0.555483	0.564793

accumulating the proportions of words occurring during the current frame's 100-ms support. The word segmentation was taken from [17]. As for syllable rate, we produced variants $\mathbf{W}^{(0.2)}$, $\mathbf{W}^{(0.3)}$, and $\mathbf{W}^{(0.4)}$. Their cross entropies for DEVSET for all supra-segmental speaking rate features explored are shown in Table 2; the numbers are similar for TRAINSET and TESTSET.

Table 2. DEVSET cross entropy rates for word rate ($\mathbf{W}$), syllable nucleus rate ($\mathbf{S}$), and their causal corrections ($\mathbf{W}_c$ and $\mathbf{S}_c$, respectively), as a function of the per-frame integration interval t_{ave}. Compare to $\mathbf{R}$ in column 4 of Table 1.

t_{ave}	$\mathbf{W}^{(t_{ave})}$	$\mathbf{S}^{(t_{ave})}$	$\mathbf{W}_c^{(t_{ave})}$	$\mathbf{S}_c^{(t_{ave})}$
0.1	0.19409	0.440787	0.555835	0.967739
0.2	0.199944	0.439226	0.542967	0.533959
0.3	0.21506	0.438476	0.540861	0.531721
0.4	0.234461	0.437888	0.541314	0.530621

As seen in column 2, speaking rate $\mathbf{W}$ obtained from word segmentation performs remarkably well, beating even $\mathbf{Q}$. The syllable-nucleus rate $\mathbf{S}$, in column 3, appears much better than $\mathbf{R}$. Further analysis, however, revealed that $\mathbf{W}$ is unfairly exploiting the future; for example, if no words were observed at frame $t-2$, and only a partial word was observed at $t-1$, it is obvious that the rest of the word must be at t. Similarly, the $\mathbf{S}$ speaking rate feature implicitly relies on a look-ahead in order to posit peaks. We attempted to correct for these "cheating" defects, by eliminating not-completed words from $\mathbf{W}$ and nuclei posited during $t-1$ for $\mathbf{S}$, leading to the cross entropies in columns 4 and 5. These appear to be broadly similar to $\mathbf{R}$ in performance, suggesting that $\mathbf{R}$ is in fact measuring speaking rate, at the segmental level, as argued in [10]. We conduct subsequent speaking rate experiments with $\mathbf{R}$ only.[2]

For completion, it should be noted that in spite of similar performance on the prediction task, the correlation among $\mathbf{R}$ and the variants of $\mathbf{W}$ and of $\mathbf{S}$ is smaller than could be expected. Over all of TRAINSET, DEVSET, and TESTSET it is only 0.64 between $\mathbf{S}_c^{(0.4)}$ and $\mathbf{W}_c^{(0.4)}$, 0.49 between $\mathbf{R}$ and $\mathbf{S}_c^{(0.4)}$, and 0.45 between $\mathbf{R}$ and $\mathbf{S}_c^{(0.4)}$. These correlations are highly statistically significant ($p < 0.0001$), however, as they are based on millions of frames.

3.4. Concatenation of Features

Finally, we compute the cross entropy rates for conditioning concatenations of energy $\mathbf{E}$ or speech rate $\mathbf{R}$ with speech/non-speech activity $\mathbf{Q}$, and compare them to those obtained using $\mathbf{Q}$ alone. The results are shown in Figure 2.

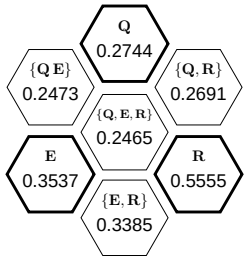

Fig. 2. DEVSET cross entropy rates (in bits per 100-ms frame) for all conditioning-context combinations of $\mathbf{Q}$, $\mathbf{E}$, and $\mathbf{R}$. Results for TESTSET are similar.

We observe that while $\mathbf{E}$ and $\mathbf{R}$ are much weaker than $\mathbf{Q}$ at 10 Hz, both contribute to prediction in combination with $\mathbf{Q}$. The effect of $\mathbf{E}$ is strong: it reduces cross entropy by $0.2744 - 0.2473 = 0.0271$ bits per 100-ms frame, or 0.271 bits/second. The effect of $\mathbf{R}$ is weaker: its reduction over $\mathbf{Q}$ alone is $0.2744 - 0.2681 = 0.0054$ bits per 100-ms frame, or 0.054 bits/second. Finally, although $\mathbf{E}$ and $\mathbf{R}$ are somewhat complimentary ($\mathbf{R}$ reduces cross entropy over $\mathbf{E}$ alone by $0.3537 - 0.3385 = 0.0152$ bits per 100-ms frame, or 0.152 bits/second), $\mathbf{R}$ is not very helpful if both $\mathbf{Q}$ and $\mathbf{E}$ are available.

4. DISCUSSION

4.1. Analysis of Global Model Surprise

To understand how energy and MSF impact the prediction task, we conducted a cursory analysis of the histogram of model "surprise" on DEVSET for models based on $\mathbf{Q}$, $\{\mathbf{Q}, \mathbf{E}\}$, or $\{\mathbf{Q}, \mathbf{R}\}$. For $\mathbf{Q}$ without either $\mathbf{E}$ or $\mathbf{R}$, the surprise of 50% of frames falls into three very narrow regions. The first of these acounts for 22.5% of frame mass, and corresponds to instants in which the target speaker has been silent and their interlocutor has been speaking, and the target speaker remains silent. Additionally conditioning on $\mathbf{E}$ or $\mathbf{R}$ redistributes the mass in a symmetric fashion around the $\mathbf{Q}$-only peak (within ± 0.001 bits per 100 ms), and thereby seems to have little net effect. In the second region, accounting for 21.3% of frames, the target speaker has been speaking and their interlocutor has been silent, and the target speaker continues to speak. In these cases, conditioning additionally on $\mathbf{E}$ reduces entropy rates by 3-4 times as much as conditioning additionally on $\mathbf{R}$ (0.0037 vs 0.0010 bits per 100 ms). Finally, in the third region, the interlocutor has been silent but the target speaker changes state; these frames account for 5.4% of the frame mass in DEVSET. Here, the impact of conditioning additionally on a prosodic feature is large in absolute terms: for $\mathbf{E}$ it is almost twice as large as for $\mathbf{R}$ (0.2497 vs 0.1370 bits per 100 ms). In the remaining 50% of frames, the role of $\mathbf{E}$ nor $\mathbf{R}$ appears to depend on the specific distribution of speech activity in the conditioning history for the two parties.

4.2. Analysis of Local Model Surprise

A temporally local analysis, that of the *evolution* of cross-entropy, is depicted for a snippet of a randomly drawn DEVSET conversation in Figure 3. The snippet is just over 10 seconds long. One conversant,

[2]We have an additional reason for not preferring $\mathbf{S}$ over $\mathbf{R}$: it uses energy peaks to infer syllable peaks. Energy is explicitly modeled in $\mathbf{E}$, and therefore exploiting $\mathbf{S}$ confounds our comparison. Some of the combinatorial experiments required to shed light on the complimentarity of all these features are currently underway.

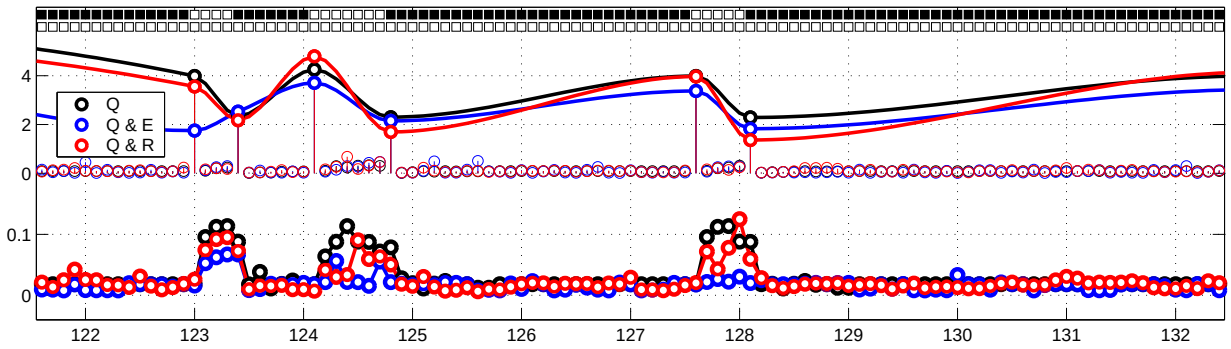

Fig. 3. Evolution in time (in seconds, along x-axis) of cross entropy ("surprise") for three models, for 10 seconds of dialogue 3161 from the DEVSET. Snippet of the two-row speech/non-speech chronogram shown at top. Surprise trajectories for the top-row (speaking) conversant shown as stem plots, with talkspurt-boundary cross entropies connected with cubic Hermite interpolations for visualization. Surprise trajectories for the bottom-row (not speaking) conversant shown at bottom; note differences in cross entropy scale (along the y axis) for the two conversants' trajectories.

shown at the top, produces 4 pause-separated talkspurts, while the other, at bottow, is silent during the whole interval shown.

The figure shows that the $\mathbf{Q}$-only model surprise for the speaking party departs considerably from zero only during changes between speech and non-speech. This is expected in the absence of interlocutor talk, since state-change frames are more rare than frames in which state does not change. For the silent party, the $\mathbf{Q}$-only model is surprised during the speaking party's pauses. Here, in the most recent portion of the conditioning history neither party is observed to be speaking, which is more rare than one party speaking.

When energy is additionally available in the conditioning context, cross entropies are lower for both the speaking and the silent parties. For example, the $\{\mathbf{Q}, \mathbf{E}\}$ model appears only half as surprised as the $\mathbf{Q}$-only model when the speaking party falls silent at 123 seconds. This suggests that the energy trajectory of the speaking party, prior to that moment, contains information that is predictive of the pause. Similarly, the $\{\mathbf{Q}, \mathbf{E}\}$ model is not at all suprised by the silent party's failure to start speaking during the speaking party's pause at 128 seconds. It suggests that the speaking party's preceding talk exhibited an energy trajectory which signalled that s/he would shortly continue speaking. When MSF is available, talkspurt termination is more surprising than when energy is available instead; MSF also has a weaker impact on the prediction that the listening party will stay silent during a speaker's pauses.

4.3. Relation to Prior Work

The STT framework presented here is the same as that in recent work on "stochastic turn-taking" (STT) models [9], where it was applied to study the impact of $\mathbf{E}$ inclusion in the conditioning history. The framework has not previously been applied to the MSF feature, which was introduced in [10] and evaluated at talkspurt ends in face-to-face conversations conducted in Swedish. Prior work on STT models has not analyzed the evolution of cross-entropy as a function of talkspurt context.

4.4. Potential Impact

We believe that the proposed framework may have considerable impact for practitioners trying to determine how prosody shapes conversations, in a data-driven fashion. Provided that a prosodic feature can be encoded as a frame-level phenomenon and attributed to individual parties, the framework requires only that a speech/non-speech segmentation be available. Under these conditions, evaluating and comparing the impact of a feature on the evolution of "surprise" is only a matter of training a neural network. It requires no orthographic transcription, no part-of-speech tagging, no manual dialog act annotation, and is generally quite theory-agnostic. We expect that it can be used "out-of-the-box" to compare prosodic practice across corpora representing different domains and languages. With some effort, the framework can also be extended to permit inclusion of categorical variables such as words and tags.

5. CONCLUSIONS

We have presented an information-theoretic framework for the analysis of the role of frame-level features in shaping the incipient deployment of speech by participants to dialogue. The framework enables a numerical comparison between arbitrary prosodic features. To illustrate this, we deliberately chose one feature (energy) whose effect is relatively well-documented as well as one on which the literature is not often in agreement (speaking rate). Our results demonstrate that the second feature is on average less predictive of incipient speech for both parties to dialogue conducted over the telephone in English. Furthermore, the presented method was shown to be capable of identifying the specific instants during which specific features are operative in mitigating surprise. Because the only annotation necessary in order to apply the method is a speech/non-speech segmentation, it is expected that similar discovery experiments can be cheaply and automatically conducted for a large number of situated speech corpora, and thus shed new light on the variability and consistency of prosodic practice in conversation.

6. ACKNOWLEDGMENTS

Anna Hjalmarsson is supported by the Swedish Research Council (VR) project *Classifying and deploying pauses for flow control in conversational systems* (2011-6152). Computing resources at Carnegie Mellon University were available courtesy of Qin Jin.

7. REFERENCES

[1] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, pp. 696–735, 1974.

[2] Antoine Raux and Maxine Eskenazi, "Optimizing endpointing thresholds using dialogue features in a spoken dialogue system," in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, 2008, pp. 1–10.

[3] Agustín Gravano and Julia Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[4] David Schlangen, "From reaction to prediction: Experiments with computational models of turn-taking," *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking*, 2006.

[5] Starkey Duncan, "Some signals and rules for taking speaking turns in conversations.," *Journal of personality and social psychology*, vol. 23, no. 2, pp. 283, 1972.

[6] John K Local, John Kelly, and William HG Wells, "Towards a phonology of conversation: turn-taking in tyneside english," *Journal of Linguistics*, vol. 22, no. 02, pp. 411–437, 1986.

[7] Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke, "A prosody-based approach to end-of-utterance detection that does not require speech recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 1, pp. I–608.

[8] Margaret Zellers, "Duration and pitch in perception of turn transition by swedish and english listeners," in *Fonetik 2014*, 2014.

[9] K. Laskowski, "Exploiting loudness dynamics in stochastic models of turn-taking," in *Proc. 4th IEEE Workshop on Spoken Language Technology*, Miami FL, USA, 2012, pp. 79–84.

[10] A. Hjalmarsson and K. Laskowski, "Measuring final lengthening for speaker-change prediction," Firenze, Italy, 2011, pp. 2065–2068.

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

[12] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Tech. Rep., Center for Research in Computing Technology, Harvard University, Cambridge MA, USA, 1998.

[13] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[14] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York NY, USA, 1991.

[15] J. J. Godfrey and E. C. Holliman, *Switchboard-1 Release 2*, vol. LDC97S62, Linguistic Data Consortium, Phildelphia PA, USA, 1997.

[16] K. Laskowski and E. Shriberg, "Corpus-independent history compression for stochastic turn-taking models," in *Proc. 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, IEEE, pp. 2189–2192.

[17] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of SWITCHBOARD," in *Proc. 5th International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, November 1998, ISCA, vol. 4, pp. 1543–1546.

[18] Nivja H de Jong and Ton Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.