From Reaction To Prediction
Experiments with Computational Models of Turn-Taking

David Schlangen

Department of Linguistics
University of Potsdam, Germany
das@ling.uni-potsdam.de

Abstract

Deciding when to take (or not to take) the turn in a conversation is an important task. It has been stressed in the descriptive literature that such decisions must involve *prediction*, as they often seem to be made before a transition place has been reached. In computational systems, however, turn-taking is normally a *reaction* to parameters like pause length. In this paper, we report on experiments that try to bridge this gap. We describe an experiment (using controlled stimuli) that shows human performance at prediction of turn-taking decisions and then show that a model automatically induced from data can reach a similar level of performance. We then describe a series of experiments on spontaneous dialogue data where we combine pause thresholds with syntactic and prosodic information to make turn-taking decisions, successively reducing the pause threshold until *reaction* becomes *prediction*. All our classifiers improve significantly over the baselines; *prediction* however is shown to be the hardest task, and we discuss additional information sources that could improve it.

Index Terms: turn-taking, machine learning, prosody.

1. Introduction

In their seminal 1974 paper, [1] made the claim that turn-taking decisions in dialogue must be based to some extent on *prediction*, which can explain both the small average gap between turns of different speakers and the systematic mistakes (overlaps) that do occasionally occur. Subsequent research has done much to identify the information sources that play a role in this decision process (see brief overview in the next section). Spoken dialogue systems, on the other hand, often for technical reasons implement a *reactive* model, where a pause threshold (typically between 0.5 and 1 second, [2]) is used to determine whether the current speaker wants to yield the turn. (Work exists that explores the use of more flexible thresholds, see next section.)

The work reported in this paper is intended to help bridge this gap. We report on experiments on computational modelling of turn-taking decisions, based on corpora of human–human dialogue. We show that relatively straightforward prosodic modelling (using f0 features and intensity) can achieve a performance on a prediction task ("does the turn continue after this utterance or not?") that is comparable to that of human subjects on the same (controlled) data. We then describe a series of experiments on a corpus of spontaneous dialogue utterances, where we move from classifying *pauses* (of decreasing lengths) to classifying all *words* into whether they end the turn or not. In all these tasks, our classifiers, using prosodic and syntactic information, improve over the baselines. In the latter, hardest task, however, the improvement is relatively smaller.

The remainder of the paper is structured as follows. In the next section, we briefly review descriptive/experimental and computational work on turn-taking; from both we take inspirations for the features which we use to represent our data. These are described in Section 3. Our experiments are reported in Section 4; we close with some conclusions and future work.

2. Related Work

As mentioned above, the model of [1] is *predictive*: according to this model, turns consist of units that have *projectable* regions (i.e., their occurrence is predictable) at which turn transitions are "relevant" (*transition relevance places*, TRPs). Subsequent research has helped to separate the contribution of different kinds of information—syntactic, prosodic and pragmatic—to the task of predicting TRPs.[1] [3] showed that listeners were able, after being played an utterance up to a (syntactically) potentially final word, to predict whether the utterance would continue. This effect correlated with prosodic features of the word, indicating that the potentially final word alone might hold enough prosodic information to make prediction of the future course of the utterance possible.

[4] propose a "filter model", where prosody and pragmatics *select* from syntactic completion points in order to determine potential TRPs. [5], refining a method introduced by [6], showed that prosody (in particular, certain intonational patterns, among them H* %) only contributes to *holding* the turn: If the speaker wants to hold the turn beyond a point of syntactic completion, she can use a certain intonational pattern to signal this. Corresponding turn-*yield* signals do not seem to be in evidence.

In contrast, most current spoken dialogue systems do not make use of prediction but rather rely on pause length as a cue for taking the turn. [7] and [8] describe methods for making such pause duration thresholds more flexible, using classifiers to judge whether a pause by the speaker should be understood as the end of her turn or not (e.g., being a hesitation instead). Both systems use task-specific features that represent state-of-understanding. Related but closer to our approach (to use only prosodic and syntactic features) are [2, 9], which use a staged array of classifiers that each are triggered separately at different pause lengths, until one decides (with a confidence over a certain threshold) that the pause-so-far marks an end of turn.[2]

[1] There has also been much work on non-linguistic means of turn-taking management (e.g. gestures); as we look at telephone-conversations only, we will not go into this.

[2] The authors call their task "end-of-utterance detection"; we prefer to

3. Corpora and Features

3.1. Corpora and Data Preparation

The principal corpus we used for our experiments was the switchboard corpus (henceforth: *swbd*; [10]) of spontaneous human–human dialogues about general topics, in English. We also collected a (much smaller) corpus (henceforth: *pcorp*) ourselves, using a "semi-conversational" setting where a speaker (we recorded three different speakers altogether) was asked to read out descriptions of one to three sentences in length (all declarative sentences), in response to which the addressee had to identify cards with pictures. This allowed us to control the utterances (all utterance types were instantiated in all possible positions, turn-internal and at turn-boundaries) while keeping the setting conversational. This corpus is in German.

Both corpora were processed in the same way to produce feature representations. From the audio we computed "raw" acoustic features (f0 and intensity measures at 10ms frames; smoothed), using the software Praat [11], this information was merged together with aligned transcriptions and POS tags to form the basis for the computation of the features we used for learning.[3] We set aside sections 3 and 4 from the switchboard corpus for training of language models (see below), and computed features for training and testing for (each word of) 20 dialogues from section 2, as follows.

3.2. Features

The prosodic modelling was relatively straightforward, and follows work such as [13]: we computed a number of features that describe the "shape" of a curve—e.g., the number of changes in direction, the tendency to go upwards or downwards, and a number of normalised measures (maximum/minimum divided by mean (of unit or general mean of speaker); standard deviation around mean; difference value at boundary to mean, by standard deviations), both for f0 curves and intensity. For f0 alone, we also segmented the unit into start, middle and end, and computed means and SD for those, to be able to characterise these regions individually. This results in 29 features in the `prosodic` group. We also additionally used the length of the word as a feature.

Syntactic properties of the data are modelled with 8 features. We computed language models of word and POS sequences, using the CMU LM toolkit [14], adding a pseudo-token *end-of-utterance* at appropriate places.[4] With these we computed for each word w_n in the data the probability of the sequence w_{n-2}, w_{n-1}, w_n, *end-of-utterance*. We also computed prior probabilities of (word and POS) types ending an utterance (again on *swbd3-4*); a fallback for tokens for which no such priors were available was the overall prior probability of a token ending an utterance (#tokens in utterance-final position / #tokens in training set). Hence, the values of the feature are *not* in a proper probability distribution. Note that these features are all conditioned on *utterances*, not *turns* – they are meant to model TRPs, at which prosody helps to predict the upcoming transition type. Lastly, we have features representing the length of the current utterance and the current turn up to the current word.

For each word, we also recorded whether in the reference transcription it was marked as being an utterance boundary, and if so, which kind of transition follows (*take*: a different speaker takes the floor; or *wait*: same speaker continues). This is the class feature to be learnt.

4. Experiments and Results

4.1. Utterance Classification Task, Human Subjects

In the first experiment we used the software *Linger*[5] to present utterances from *pcorp* to human subjects (24 university students with no reported hearing difficulties) who were asked to judge whether the speaker of the utterance would continue speaking afterwards (i.e., hearer should *wait*) or not (i.e., hearer can *take*). Each condition was equally likely in the presentations. Figure 1 shows the results, as success rate (overall and by condition); we also give the f-measure for each class.[6]

The way the task was set up ensured that the human subjects did not have more than acoustic information to base their decision on: all utterances were complete sentences (syntactically and hence semantically 'complete'), and as there was no surrounding context, there was no basis to tell whether the presented utterance was a pragmatically complete contribution in itself or not. As reported in similar studies in the literature (see above), the subjects performed better at recognising "keep turn" situations (class *wait*).

class	cor.	incor.	f-m
overall	56%	44%	–
wait	78%	22%	0.63
take	34%	66%	0.38

Figure 1: Results for Exp. 1

4.2. ML on Utterance Classification Task

We then used machine learning (ML) on the same corpus (this time the full set, 385 data points).[7] The decision to be made is the same (*wait* vs. *take*, at known utterance boundaries), however, the utterances are represented here by (acoustic features from) their last word only, following observations from the literature (cited above) that acoustic information used in this task is concentrated on that last word. No other information sources were available to the learners, just as in the setting with human subjects.

Results are shown in Table 1. The one-rule learner already improves significantly (paired t-test) over the majority baseline.

keep the distinction between "utterance" (as realisation of a unit roughly comparable to the syntactic sentence) and "turn" (which may consist of several utterances).

[3] Full word-alignment and POS tags were available only for *swbd*; the word-alignment data was provided by the SWBD-project at the MS state university, POS-tag information and utterance and turn segmentation information is available as part of the Penn-treebank distribution ([12].

[4] These syntactic features were only computed for *swbd*.

[5] `http://tedlab.mit.edu/~dr/Linger`

[6] A single sample t-test shows that the overall performance is significantly better than chance ($p < 0.05$); a three-way ANOVA shows there was *no* significant difference between recordings from different *speakers*; the difference between *wait* and *take* was highly significant ($p < 0.005$; t-test).

[7] In all following ML experiments, we used various machine learning algorithms (Ripper, C4.8, Bayesian Networks) as implemented by the WEKA toolkit [15]. We refer the reader to [15] for references to the original work describing these classifiers.

If not stated otherwise, all results were obtained by performing 10-fold cross-validation. For feature selection, we used *ClassifierSubsetEqual* with *greedy forward search*, performed on a hold-out data-set that was not used for evaluating the classifier. For reasons of space, we report only results for the best-performing (or otherwise relevant) classifiers.

As baseline we used majority class prediction and a one-rule decision tree learner (known to perform, despite its simplicity, often rather well).

Clsf	FSet	C%	IC%	Cls	F
Maj.	full	53.2	46.7	wt	0.69
				tk	0
OneR		57.4	42.6	wt	0.62
	fml3	(w-f: 0.57)		tk	0.52
JRip		**68.0**	**31.9**	wt	**0.72**
	fcdr,fml1,ipxm	(w-f: 0.68)		tk	**0.63**
JRip		65.2	34.8	wt	0.69
	f0 only			tk	0.59
JRip		60.8	39.2	wt	0.62
	int only			tk	0.59

Table 1: Results Experiment 2. Clsf = Classifier Used; FSet = Features available; (I)C% = Percent (in)correct; Cls = Class; F = F-measure

It chooses fml3 (the normalised f0-mean in the last third of the word). The best performing classifier shows a further relative improvement of 19% (on class-size weighted average of f-measures), while relying on just three features, two f0-based ones (number of direction changes in curve and normalised mean in the *first* third) and one intensity feature (distance of the maximal intensity from the mean intensity). For reasons of space, we cannot discuss the learnt rules in detail; roughly, they seem to correspond to observations from the descriptive literature (e.g., low intensity and 'trailing off' at end is learnt as a characteristic of end of turns; a combination of features that approximates flat mid-level tones as a characteristic of *wait*).

The results for reduced features sets (f0 only and intensity only; shown in the next rows in the table) suggest that *take* relies more on a combination of f0 and int values, whereas *wait*-information is localised more on f0 patterns. (This is also in line with the literature discussed above.) Overall, compared with human performance on the same corpus, our classifiers show the same tendency to better predict the wait class (which however in this data, unlike in the stimuli presented to the subjects, is the majority class by a slight margin).

Table 2 shows the results of the best classifier when the same experiment (same classes to be learnt, same features) is performed on data from *swbd*, which presents several additional challenges: First, the *technical* quality varies more compared to our own corpus (occasional background noise, distortions, crosstalk, etc.). Second, and more importantly, the data itself is more varied, containing of course not only declarative sentences but all kinds of utterances that can be expected in spontaneous conversation; among others, *back-channel utterances* (BC; "uhu", "yeah") which are interesting for our discussion because they are generally considered as not instantiating turn-*transitions*, despite representing speaker changes. We followed this line (as for example [6] does as well) and treated utterance boundaries followed only by BCs as *wait*.[8] The results are slightly worse compared to Experiment 2, with the same pattern of *wait* being predicted better than *take* (but given an even stronger imbalance btw. classes).

We conclude from these experiments that upcoming transitions, given information about the acoustic shape of the final word (and hence, given knowledge that the current word is *utterance-final*), can be modelled computationally with a performance comparable to that of human subjects. We now move to the (more complicated) task of additionally recognising whether a current word is utterance-final or not.

[8] The difference in *language* btw. the corpora should not be of significance: at least w.r.t. the respective role of intonation in turn-taking, English seems to behave similarly to German [5]; but we leave further study of this interesting question to future work.

Clsf	FSet	C%	IC%	Cls	F
BayN	full	64.6	35.4	wt	0.74
		(w-f: **0.65**)		tk	0.46

Table 2: Classifying Utterance Boundaries on *swbd* data

PTR	clsf	f-w	f-t	FAR	
.500s	bsln	0	62.8	54.2	
	J48	71.8	65.3	33.2	$\Delta = 38.7\%$
.250s	bsln	0	50.3	66.3	
	J48	76.7	50.5	46.9	$\Delta = 29.3\%$
.100s	bsln	0	41.7	73.6	
	J48	82.3	44.1	51.3	$\Delta = 30.3\%$
0s	bsln	n/a	n/a	n/a	
	J48	97.6	35.5	41.6	(see Tab. 4)

Table 3: Results at various pause thresholds (PTRs)

4.3. Detecting Turn-Endings using Pause Thresholds

The now frame the task in a way that is more similar to the problem practical systems face. Again we trained classifiers on features of a potentially turn-final word; however, instead of relying on the human annotation to identify the set of these words, we selected them according to a "length of following pause" criterion. With this data-set, the task becomes one of distinguishing between hesitations within a turn and turn-ends. By reducing the pause length threshold (and hence increasing the number of candidate words), we then move this task closer again to the theoretically motivated one of modelling *prediction*.

Table 3 shows the results (of the best classifier) at various pause threshold (PTR) settings (f-w/f-t: f-measure for *wait* and *take*, respectively; FAR: false alarm rate (for take), where FAR = FP / (TP + FP). As a baseline we use a strategy that is often used in practical dialogue systems (see discussion above), where *all* pauses are classified as *take*.[9] As the table shows, our classifiers can in all cases improve on the baseline (i.e., they block certain pauses from being classified as marking *end-of-turn*). Interestingly, as the PTR is shortened (and hence the size of the data-set and its imbalance increases), the improvement decreases (the FAR increases even stronger for the classifiers). We explore this more with the final experiment, where the PTR is set to 0.

4.4. All Word Classification Task

In this task, *all* words are classified, which means that besides the words contained in the previous data-sets (words at hand-coded utterances boundaries and before pauses) this one contains also "intra-utterance" words. It hence is massively (1:20) biased towards *wait*. Table 4 gives the results. FAR is not as useful as a metric here, as it only looks at false positives, and so would be triv-

[9] As a side remark, we note that our corpus seems to contain much more non-turn end pauses than the corpus used by [9] (human to computer utterances in a task-oriented domain), as here the baseline performs much worse.

Clsf.	Fl.	FSet	C%	IC%	Cls	F
Maj.	–	full	95.6	4.37	wt	0.98
(κ: 0; TT_e: undef.)			(w-f: 0.93)		tk	0
OneR	–	priw	95.4	4.6	wt	0.98
(κ: 22; TT_e: 6.96)			(w-f: 0.95)		tk	0.24
J48	–	full	95.4	4.6	wt	0.98
(κ: 33; TT_e: 4.81)			(w-f: 0.95)		tk	0.36
J48	–	syn.	95.8	4.2	wt	0.98
(κ: 24; TT_e: 6.53)			(w-f: 0.95)		tk	0.26
J48	–	ac.	95.4	4.6	wt	0.98
(κ: 6; TT_e: 31)			(w-f: 0.94)		tk	0.07

Table 4: All Word Classification Task

ially zero for a classifier that only chooses the other class; to get a more balanced picture (while still punishing false alarms more than missed opportunities to *take*), we give an *ad hoc* metric TT_e defined as $(\text{FP} * 2 + \text{FN})/\text{TP}$. We also show κ, which captures how much has been learned by discounting what could have been correct by chance.

The one-rule learner (which chooses prior word prob.) already can improve significantly over the majority class baseline, and the best-performing full classifier further improves on this. Testing the blocks of acoustic and syntactic features individually shows that while syntax has the biggest contribution, acoustic information is clearly relevant here. However, the overall performance of the classifiers on this, the most difficult of the tasks discussed here, is not very good, and recall of *take* decisions is low. While from the experimental work cited above it can be expected that without *pragmatic* information and with (such simple) syntax and prosody alone performance will be limited, future work will have to show if more high-level, sophisticated modelling can improve here.

5. Conclusions and Future Work

We have presented several experiments on modelling turn taking decisions, using various data-sets with different criteria for selecting the candidate token, and moving from *prediction* on controlled data to *reaction* (to "silence events" / pauses) on spontaneous data back to *prediction* on the latter data. Our (f0 and intensity-based) prosodic and (n-gram based) syntactic features were shown to be of use in these tasks.

Some of the features we've used could be made available in real-time for a practical system; for others we relied on reference annotations. As one direction for further work, we want to explore what the influence of lower quality input would be; for this, we will simulate ASR output by introducing controlled word errors. We will also systematically explore the use of features that "lag behind", i.e. only represent the data up to a certain time before the point at which the decision is to be made. On the theoretical side, we want to explore the contribution of syntactic information more, and will integrate higher-level information (phrase-boundaries and chunk parses).

Acknowledgements: Thanks to Volker Strom for some initial help with the modelling, to the Potsdam Discourse Group (esp. Katja Jasinskaya) for some discussion, and to the anonymous reviewers for their helpful comments.

6. References

[1] Harvey Sacks, Emanuel A. Schegloff, and Gail A. Jefferson, "A simplest systematic for the organization of turn-taking in conversation," *Language*, vol. 50, pp. 735–996, 1974.

[2] Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke, "Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody," in *Proceedings of ICSLP2002*, Denver, USA, September 2002.

[3] François Grosjean, "How long is the sentence? Prediction and prosody in the on-line processing of language," *Linguistics*, vol. 21, pp. 501–529, 1983.

[4] Cecilia E. Ford and Sandra A. Thompson, "Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns," in *Interaction and Grammar*, E. Ochs, E.A. Schegloff, and S.A. Thompson, Eds., pp. 134–184. CUP, Cambridge, UK, 1996.

[5] Johanneke Caspers, "Local speech melody as a limiting factor in the turn-taking system in dutch," *Journal of Phonetics*, vol. 31, pp. 251–276, 2003.

[6] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map-task dialogs," *Language and Speech*, vol. 41, no. 3–4, pp. 295–321, 1998.

[7] Linda Bell, Johan Boye, and Joakim Gustafson, "Real-time handling of fragmented utterances," in *Proceedings of the NAACL-01)*, Pittsburgh, USA, June 2001.

[8] Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyoaki Aikawa, "Learning decision trees to determine turn-taking by spoken dialogue systems," in *Proceedings of ICSLP-2002*, Denver, USA, September 2002, pp. 861–864.

[9] Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke, "A prosody-based approach to end-of-utterance detection that does not require speech recognition," in *Proceedings of ICASSP2003*, Hong Kong, China, 2003.

[10] John J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and devlopment," in *Proceedings of ICASSP-1992*, San Francisco, USA, March 1992, pp. 517–520.

[11] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9–10, pp. 341–345, 2001.

[12] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz, "Building a large annotated corpus of english: the penn treebank," *Computational Linguistics*, vol. 19, pp. 313–330, 1993.

[13] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1–2, pp. 127–154, 2000.

[14] Philip R. Clarkson and Roni Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit," in *Proceedings Eurospeech 1997*, 1997.

[15] Ian H. Witten and Eibe Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, USA, 2nd edition, 2005.