Dialog Prediction for a General Model of Turn-Taking

Nigel G. Ward, Olac Fuentes, Alejandro Vega

Department of Computer Science, University of Texas at El Paso, USA

nigelward@acm.org, ofuentes@utep.edu, avega5@miners.utep.edu

Abstract

Today there are solutions for some specific turn-taking problems, but no general model. We show how turn-taking can be reduced to two more general problems, prediction and selection. We also discuss the value of predicting not only future speech/silence but also prosodic features, thereby handing not only turn-taking but "turn-shaping". To illustrate how such predictions can be made, we trained a neural network predictor. This was adequate to support some specific turn-taking decisions and was modestly accurate overall.

Index Terms: dialog system, predictive, prosody, endpointing, back-channeling, time, interaction control, dialog model

1. Introduction

Current dialog systems generally have rigid turn-taking, where the user and the system do not overlap and turn exchanges are marked unambiguously by longish periods of silence. Human-human dialog is not like this: turn-taking is swifter and more flexible, thanks largely to the use of prosodic indications of turn-taking intentions. The potential benefits of better turn-taking include improved efficiency (less dead time), higher accuracy (due to better selection of the portions of the signal to feed to the speech recognizer), and reduced cognitive load (due to fewer unexpected or disruptive system behaviors).

Recent advances in incremental speech recognition and in the modeling of turn-taking cues have enabled the demonstration of impressive abilities addressing some specific issues in turn-taking, notably endpointing [1], back-channeling [2], filler production [3], and whether to produce interleaved acknowledgments [4]. However it could be advantageous to have a general model of turn-taking, subsuming all such specific models. This paper presents such a general model and some initial steps towards its realization.

2. Predicting to Support Turn-Taking

Good dancing depends on the ability to predict the future actions of the partner: not only what they are going to do, but also when they will do it, how much energy they will put into it, how large the motion will be, and how long it will last. Good dialog is similar. The role of prediction in dialog in general and its use for specific aspects of turn-taking has been discussed before [1, 2, 5], but in order to support good turn-taking via a general model, we believe that a predictor needs certain properties.

First, a predictor should **make predictions constantly**, rather than only at certain points — such as user pause points or other timepoints meeting certain criteria — for the sake of swiftness of response.

Second, it should **predict some distance out into the future**, as suggested by Figure 1, rather than only making a prediction for one instant. One reason is to improve responsiveness for real systems, which need look-ahead since it takes time to choose and prepare responses [6]. Another reason is to support non-binary turn-taking decisions. This could be useful, for example, in a situation where the system has prompted the user for his account number and he has produced a few digits, and where the options may include "stay silent and wait", "produce a brief acknowledgment", "produce a filler while the recognizer does its job and then repeat back the first digit group", or "re-prompt". The decision among these options is, in part, a turn-taking decision. If there is a prediction that the interlocutor will continue talking for another half second, pause briefly and then resume talking, then the system has clear guidance as to which of its options fits best. Thus the use of a general predictor can support selection among more options for system-side action. This may not have much value for today's system, which at any given point usually have only one or two actions available, but as dialog managers become more powerful, and as speech synthesis becomes more flexible, this will become more useful.

Third, as hinted above, a predictor should directly **predict the system's own actions**. If we have a corpus of dialogs, where we have access to both sides, predicting the future actions of one speaker is no harder than predicting the future actions of the other. Thus we can predict the actions of both interlocutors, not only those of the user, but those of the system itself. We can make such predictions of system behavior come true by having the system behave so as to fulfill those predictions. If these predictions are corpus-based, then the system will behave in accordance with the patterns of behavior common in the corpus. This is somewhat novel, in that previous predictors generally predict only the user's actions, requiring some additional reasoning to determine how the system should behave in response. Instead predicting the system's actions directly gives a clear role for the knowledge involved in turn-taking, as seen in Figure 2 (rather than leaving it implicit, for example, in the algorithms and parameters of an endpointer). This approach requires an additional module, the Selector in the figure, to combine the information from the predictor and the content-generating components, but this can be fairly simple. For example, a basic selector might do no more than chose, among the options provided by the dialog manager, the one that best matches the prediction, and thus conforms best to turn-taking norms.

In general this framework requires that the **predictions are**

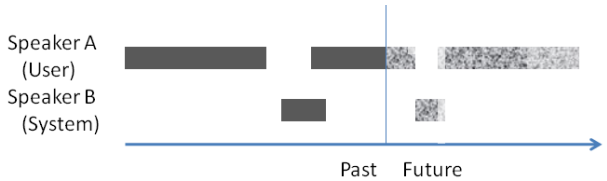

Figure 1: Predicting who will be speaking out beyond the next instant

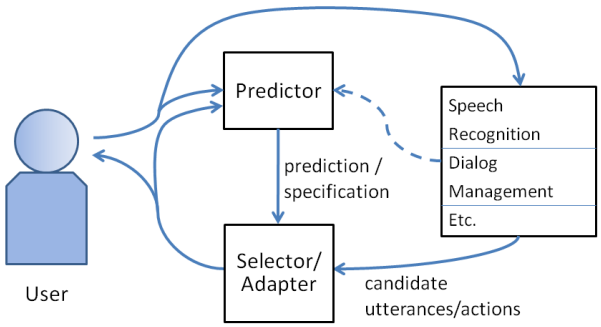

Figure 2: Separating out a predictor

probabilistic, or at least numeric, rather than absolute. One reason is to enable use of information about the strength of turn-taking cues [7], rather than just their presence or absense. Another is to support combination of turn-taking predictions with those based on other considerations, such as the content or impression [8] to convey, in order to determine which action is best overall. (Although we do not rule out the possibility that sometimes the predictions alone might determine the system's action: for example, if the user clearly wants to hold the floor, the system generally ought to remain silent.) This decision-making style, although novel for dialog systems, resembles mainstream practice in statistical machine translation, where the possible outputs are scored both by the translation model (according to their faithfulness to the source-language content), and by the language model (according to their suitability for the target language), meaning that the translation ultimately chosen is one which scores highly on both counts. Similarly, a predictor can function as a "dialog model" which, like the language model, provides quantitative judgments of the suitability of various possible responses. This makes turn-taking considerations explicit and numeric, rather than hidden in the system's control structure, as seen when an all-controlling endpointer invokes and gates the other modules.

This framework can handle the classic problems of turn-taking. Consider, for example, the problem of deciding when the other person has finished speaking (the endpointing problem). If the prediction is that the system will be speaking over the time period starting 50 milliseconds from now and ending 500 milliseconds later, then this is enough to guide the system to behave properly, without the need for a special-purpose module for endpointing. Similarly for production of a back-channel; the same mechanism will work, without the need for a special-purpose reactive component. This framework generalizes to other patterns, beyond just the handful of turn-taking patterns with specific names; thus supporting more flexible-turn-taking.

3. Beyond Turn-Taking: Turn-Shaping

If we expand the role of the predictor slightly, having it also **predict aspects of the upcoming prosody**, we can handle some related choices about what to say next. For example, if a predictor indicates that upcoming there should be a short region of speech in which pitch height, pitch range, energy, and rate are all low, then this is in effect a specification for a back-channel response. A prediction for a region of speech with initially high pitch and energy and a slow rate, followed by lower energy and less extreme values for the other features would be in effect a prediction for a turn grab.

Thus a set of predictions that includes values for key prosodic features can indicate what to do when the possible actions include, for example, a back-channel, an explicit acknowledgment, a short follow-up question, a filler followed by an acknowledgment, and so on. Making such choices can be called a process of determining "turn shape". Today's systems do this implicitly as part of content selection, generation or synthesis, but this framework handles it in the same way as turn-taking. The potential advantage is more flexibility and better responsiveness, and thus better flowing dialogs.

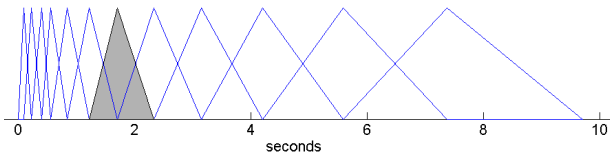

Figure 3: The 12 prediction regions, with one highlighted

4. A Predictive Model

As a proof of concept, we set out to build a predictor.

4.1. Output Features

It probably is not important to predict things with great precision, so we make predictions over regions of time. There should be more resolution in the immediate future and less in the more distant future, for which there is more opportunity to revise things as more information comes in. We thus predict values for 12 regions, the first one 230ms wide starting at 0ms (the point of prediction itself), and with each subsequent region 1.3 times wider than its predecessor, all overlapping (Figure 3).The windows are triangular, and convolved with the values over time to obtain the summary feature for each region. (These regions are probably adequate for modeling fairly sedate dialogs, but not for, e.g., heated discussions, where speakers may grab turns aggressively or complete each others' sentences, which would require finer resolution and thus more narrow early regions.)

Thus the task of our predictor is to predict the values for these 12 regions, for each feature of interest. The first feature to predict indicates whether the system should be speaking or not. Since the regions are not infinitesimal, what we actually want to predict is the "speaking fraction", that is, the fraction of time during that region that should be devoted to speaking. If there is no speech then there is no point in predicting anything else, otherwise we also want to predict key prosodic features, currently only average pitch, but in future also speaking rate, energy, and pitch range at least.

4.2. Input Features

As predictive features, for this first attempt, we chose to use only simple features, mainly prosodic ones. The first set was chosen to capture the recent turn shape, for example, who's been speaking the most and how engaged both speakers have been. For this, we use 5 features: speaking fraction, pitch height, speaking rate, energy, and pitch range, computed, of course, over regions prior to the point of prediction. We computed these over regions, specifically triangular regions of the same widths as before, but reflected about the point of prediction so that the higher resolution is immediately prior to this point. Pitch range is, however, computed over rectangular regions. All are speaker-normalized. There were 60 such "coarse" features for the speaker and another 60 for the interlocutor.

The second set of features was chosen to capture things

	prediction performance	
	speaking fraction	pitch average
baseline	.273	.251
self only (coarse+)	.218	.233
other only (coarse+)	.240	.239
both (coarse+)	.214	.237
both (coarse+fine+)	.216	.234

Table 1: Mean absolute error using different sets of predictive features. Time into dialog was included in all predictive sets

common in turn-taking signals, including lengthened vowels, syllables that are quieter or louder than average, and pitch downslopes, upslopes, and flat regions. These were computed over narrower windows, each 100ms wide, offset by 50ms, and computed only in the region where we expect to see turn-taking cues, namely the range from –1500 ms up to the prediction point. The specific features were the same 5 used for the coarse region plus highest delta pitch and lowest delta pitch. Including features for both speakers, there were 29 * 7 * 2 = 406 such "fine" features.

Time-into-dialog was the final predictive feature.

4.3. Implementation and Training

One advantage of this framework is the simplicity of development: since the predictor has a straightforward task, its performance is easy to evaluate, which makes training simpler. Specifically, our task is to predict the values of 24 features: the values for both speaking fraction and average pitch in each of 12 future windows. We use mean absolute error as the evaluation metric.

The data was taken from Switchboard corpus dialogs with good channel separation. The training data was 8 dialogs, about 40 minutes in all. Predictions are made every 100ms, for both tracks, giving about 48K input-output training sets. Most of the output regions are mostly silence, so only about half of the sets, those with at least two valid pitch points, were used for training the average-pitch predictor.

As a first attempt, we chose to use neural networks. The results reported below were obtained with a net using 4 hidden nodes to predict the 12 speaking fraction features, and 12 individual nets to predict each pitch average feature.

4.4. Performance

Our hypothesis is that a predictor trained on general dialog data will be able to learn specific turn-taking behaviors, given only general training. At this point the performance is not very good, but visual inspection of the predictions reveals instances of success at back-channel prediction, endpointing with fast responses, and floor holding.

Although the predictions are not good enough to be useful yet, we can see how such predictions might in future be used by considering an example. Figure 4 shows the predictions made at 35 seconds into dialog sw2375. To avoid misinterpretation, we note that, although there are pitch predictions throughout, these are without meaning when there is no speech, for example around 36 seconds. We also stress that the predictions are not calling for a specific contour, but for a set of actions which, after convolving, have roughly the desired average values for each region. Treating these predictions as a specification, for the second (bottom) track a roughly conforming action plan would be to be continue speaking continuously, inserting more pauses as time goes on, and with a small rise in pitch on the first or second word. Although this action plan is possibly acceptable, what actually happened next in the corpus was somewhat different: the speaker made a brief disfluent pause and then recovered, restarting at a higher pitch around 37 seconds. More predictions can be seen in the short video clip at http://cs.utep.edu/nigel/abstracts/chiba.html

We also measured the overall prediction accuracy, using 4 dialogs as the test set. For the evaluation of pitch predictions the error is only computed when meaningful, specifically, at times where the corpus had at least two valid pitchpoints in the region. To avoid the effects of random seeds, we trained each network 4 times and took the average performance. The baseline is the crude one of always predicting the average value in the training data. Table 1 shows the results. Overall, the predictions were only modestly above baseline. It also seems that the speaker features are more informative than the interlocutor features, and that adding more features doesn't necessarily help, possibly because of the relatively small size of the training data.

5. Future Work

This paper has presented a new general model for handling turn-taking in dialog, explained the potential advantages, described the qualities needed in a predictor to support this, designed a suitable set of prosodic features, presented a proof-of-concept implementation of the key component, and reported initial results. This suggests that even simple prediction techniques may be adequate to support more general turn-taking control.

Much work remains. First we plan to improve the predictor, by using more data, improved features or more features, including lexical or dialog state information, as suggested by the dotted arrow in Figure 2. We could also improve the modeling, perhaps by using Principal Components Analysis on the predictive features or with Conditional Random Fields.

After that we plan to create a selector/adapter, to handle the job of coming up with an utterance that matches the predictions as closely as possible. This may not be difficult: an initial selector could operate by simply choosing among fixed content (pre-recorded system prompts) based on simple matching to the predictions. A more sophisticated selector/adaptor could in addition warp a prompt to match the predicted prosodic properties, or do even more flexible generation and synthesis, while organizing [9], monitoring and adjusting its plans. The improved predictor and a simple selector will together constitute a reactive system whose performance can be evaluated.

Beyond turn-taking and turn-shaping, a dialog model of this form could be more generally useful. Insofar as emotional coloring and other aspects of prosody can be chosen, in part, by considering by the local prosodic context [10], the predictive models developed here could support such decisions. Whereas previous work on these, and other forms of accommodation and adaptation, has operated at the turn level, a predictor-based system could enable adaptations to be swifter and more precise.

Acknowledgments: We thank Gary Beverungen for programming, analysis, and discussion, and the NSF for support, as Project IIS-0914868.

6. References

[1] A. Raux and M. Eskenazi, "A finite-state turn-taking model for spoken dialog systems," in *NAACL HLT*, 2009.

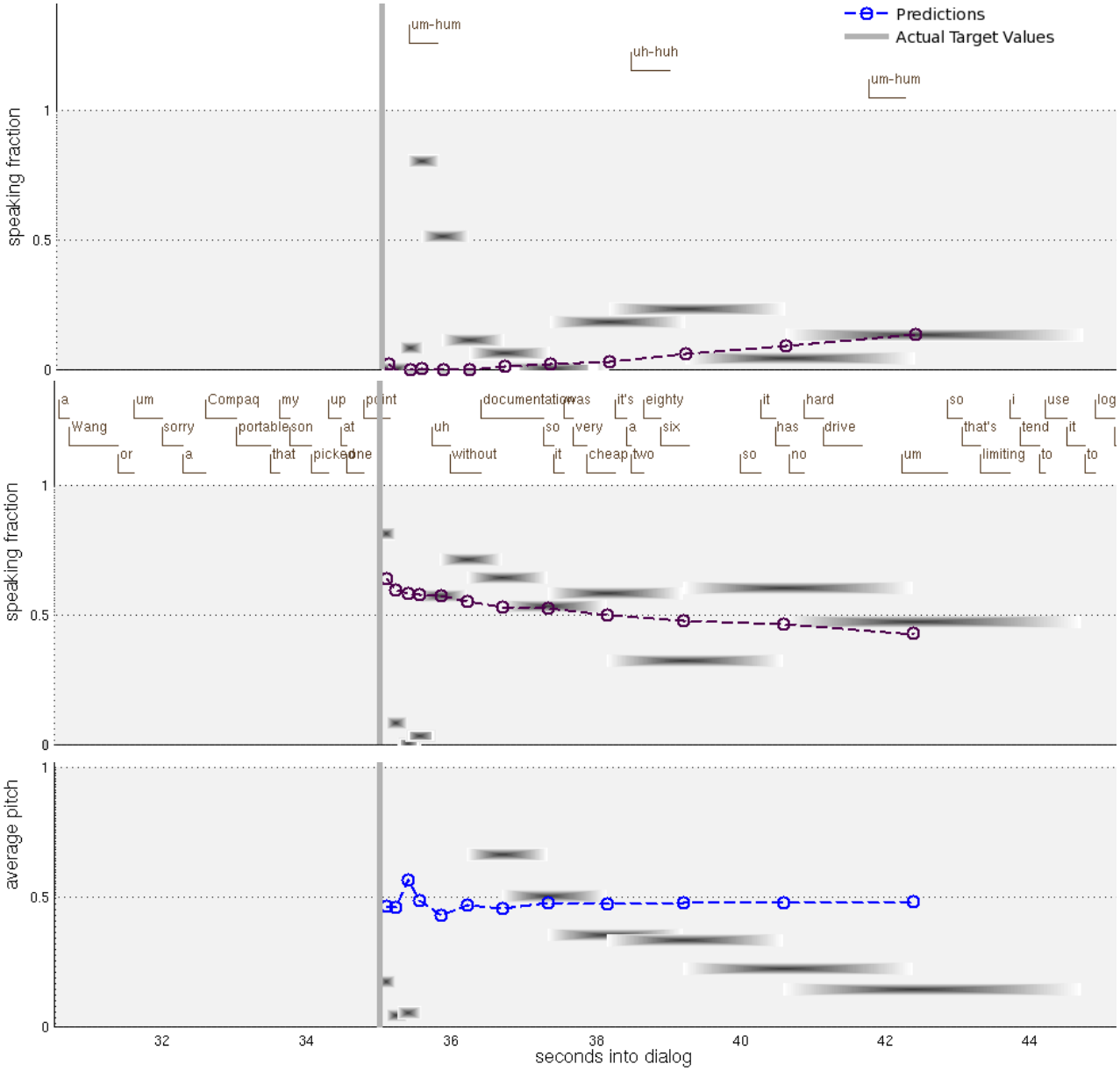

Figure 4: Example of predictions. In each case the circle-dashed line shows the predictions and the horizontal smears the targets, that is, the actual values. (Both targets and predictions are point values; the smearing and lines are shown just to make the values more visual.) The top panel is the speaking fraction for speaker 1; the bottom two panels are for speaker 2, showing showing speaking fraction above and pitch average below. The pitch values are in terms of percentile pitch for the current speaker. In each case the predictions are based on the information in the past, previous to the 35 second point (vertical bar).

[2] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, pp. 70–84, 2010.

[3] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "How quickly should a communication robot respond?," *International Journal of Social Robotics*, vol. 1, pp. 153–160, 2009.

[4] G. Skantze and D. Schlangen, "Incremental dialogue processing in a micro-domain," in *EACL*, pp. 745–753, 2009.

[5] C.-C. Lee and S. Narayanan, "Predicting interruptions in dyadic spoken interactions," in *IEEE ICASSP*, 2010.

[6] T. Baumann, "Simulating spoken dialogue with a focus on realistic turn-taking," in *Proc. 13th ESSLLI Student Session*, 2008.

[7] L. Huang, L.-P. Morency, and J. Gratch, "Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior," in *9th Int'l Conf. on Autonomous Agents and Multi-Agent Systems*, 2010.

[8] M. ter Maat and D. Heylen, "Turn management or impression management," in *Intelligent Virtual Agents 2009; LNAI 5773*, pp. 467–473, 2009.

[9] D. Bullock, "Adaptive neural models of queueing and timing in fluent action," *Trends in Cognitive Sciences*, vol. 8, pp. 426–433, 2004.

[10] J. C. Acosta and N. G. Ward, "Achieving rapport with turn-by-turn, user-responsive emotional coloring," *Speech Communication, submitted*, 2010.