Predicting the Prosody of Dialog Markers from the Prosody of the Local Context

Anindita Nath, Nigel Ward

University of Texas at El Paso
anath@miners.utep.edu, nigelward@acm.org

Abstract

Dialog systems today often use unnatural prosody. Thinking that this can be substantially alleviated by simply choosing prosody appropriate for the local context, without deep modeling of intentions or structure, we tried predicting the prosody of dialog markers from a 72 feature representation of the local prosodic contexts. This significantly decreased prediction error for the prosody of dialog markers in the Switchboard corpus, reducing the pitch error rate, for example, by 22%.

Index Terms: discourse markers, fillers, backchannels, pitch, energy

1. Motivation

Dialog systems often exhibit awkward prosody. One cause is intrinsically unnatural prosody, but this is being alleviated by recent advances leading to models that can produce utterances with natural-sounding prosody [1, 2], at least when judged in isolation. A second cause is prosody inappropriate for the intended meaning or function, and this topic has also been addressed by much work, leading to a good understanding of many prosody-meaning mappings. A third cause is prosody that is simply inappropriate for the context. This has received much less attention, except for the special case of aligning to the prosody of the interlocutor’s previous utterance, yet producing context-appropriate prosody is a major challenge for speech synthesis [3]. Here we investigate the possibility of predicting prosody directly from the prosody of the context.

Specifically, we examine the predictability of dialog markers, by which we mean discourse markers occurring in spoken dialog. These are convenient for an initial exploration because they are very common; because they serve many important functions, including managing turn-taking, marking topic structure, and expressing stance [4, 5]; because they typically are semantically semi-independent, standing outside the propositional content, or in other words, have a core procedural and not conceptual meaning [5]; and because their prosody is usually their own, being less often affected by the larger prosodic patterns that govern many word sequences.

Our hypothesis is that local prosodic context is informative for predicting the prosodic form of dialog makers.

While this has not been previously studied, other aspects of the prosody of dialog markers have received significant attention. Much work has described the prosodic correlates of various different uses, for example, discourse vs sentential uses of now [6], direction vs acknowledgment uses of right [7], questioning vs reacting uses of really [8], different polarities and intensities of yeah [9], backchannel, topic shift, and agreement marker and other uses of okay [10, 11], and so on [12, 13, 14, 15, 16, 17, 18]. Other work has noted general prosody-function mappings present across many dialog markers [19]. The study of how prosodic context directly affects dialog marker prosody has been very limited; we know of only two small-corpus studies leading to a sets of handcrafted rules [20, 21] and investigations of the extent of entrainment [22, 23].

Considering more generally work aimed at implementing simple, direct responsiveness, based on local prosodic context, there is a large body of work showing success for turn-taking predictions [24], and work on choosing the emotional coloring of responses [25] and the form of backchannels [26, 27]. This paper extends this line of inquiry to explore the prospects for prosodic tailoring of dialog makers.
log energy: mean and max

cepstral flux: mean and max

pitch: mean and max

harmonic ratio: mean and max

Figure 1: Predicted Features

2. Data

We used the Switchboard corpus of American English telephone conversations [28]. After excluding recordings with poor audio quality or artifacts that bothered our pitch tracker, we considered 1900+ conversations involving 400+ speakers. We considered all audio spans bearing labels from the list in Table 1, according to the Picone transcriptions [29]. We did not use more subjective labels [30] or do any additional checks, so cases where the word was not actually being used as dialog marker were not excluded.

3. Prosodic Features Predicted

Ultimately, we would like to predict every detail of the prosody of each dialog marker token: the value for every feature at every frame of the token. However, for this study we predict only four features namely: i) loudness, as measured by its acoustic correlate log energy, ii) pitch, as measured by its acoustic correlate fundamental frequency or f0, estimated by the pitch tracker fxrapt [31] in VoiceBox toolkit of MATLAB, iii) cepstral flux, as a measure of lengthening and reduction, and iv) the harmonic ratio [32] which is a proxy for harmonicity and, indirectly, other properties of voicing, including creakiness, breathiness, and devoicing. The relevance of pitch, energy, and timing is well known; we also included harmonicity since it appears to help differentiate among roles for many dialog markers [11]. For each of these four features, we did two experiments, one to predict the average over the entire token, and another to predict the maximum value, as both contribute to what is perceived. Figure 1 summarizes. For pitch, frames with undefined values were excluded from the final computations.

4. Context Features

As our aim is to explore, we sought neither a maximal set of features nor a minimal one. Rather we chose a set of 72 features that were diverse, convenient, reliable, and broadly covered the local context. We used contextual features for both speakers: the one who produced the dialog marker and the interlocutor. Together these features cover the time from 3.5 to 0 seconds before the start of the dialog marker and the time from 0 to 3.5 seconds after its end. We chose to consider also future information because we observed [11] that often the prosody of the dialog marker is suitable not only based on what came before, but also for what is upcoming, either by the same speaker, or by the interlocutor, to the extent the prosody of a dialog marker can guide the interlocutor’s future behavior. However, we also did experiments using only on the past context, since that may be more realistic in some scenarios. Specifically, for each speaker, we computed the 36 features shown in Table 2: 9 base features, each computed over 4 timespans. All were computed using the Midlevel Prosodic Features Toolkit [33]. The four pitch configuration features are used to enable every-meaningful computation of pitch information, even over windows with or no pitch points [34]. The “peak disalignment” feature is a measure of the displacement between energy peaks and pitch peaks [35]; for this data, this generally measures late peak (delayed peak) occurring in stressed syllables. The specific time windows were chosen based on some initial intuitions about the rate of local prosodic change relative to broader movements, and were not subsequently optimized or revisited. Together these features capture much about the local prosody and the local turn-taking state. Both the token features and the context features were z-normalized per track, to reduce the effects of intrinsic speaker differences.

5. Prediction Models

Our goal being insight rather than optimization, we used a very simple model: multivariate linear regression. This allowed us to trivially examine how the context features were affecting the dialog markers’ prosody. We developed a separate model for each dialog marker category, as we do not expect the same rules to work well for all: for both huh and okay, for example.

6. Experiment Design and Results

We followed an intra-corpus evaluation approach. Each model, one for each of the 12 dialog markers, was evaluated with a disjoint train-test split of 70:30, chosen such that the test set contained no dialogs seen in the training
set. The root mean squared error (RMSE) was used to evaluate the performance of each model. The utility of local context information was measured by the percent reduction in RMSE values for model predictions compared to the baseline of simply predicting the average over all instances of that type, for example, predicting the global average yes.

Table 2 shows the quality for the baseline predictions and the model’s predictions. The errors are lower with the model, with reductions ranging from 22% to 37%, showing that the local context is informative. The benefit is statistically significant, for all 4 predicted features in each case (matched pairs t-tests, \( p < 0.01 \)).

We also see that the reductions were greater for the maximum features than for the mean features, although this difference may be largely due to outliers.

Predictions based on features from the past context only also gave some benefit, as seen in Table 4 with RMSE reductions of 0.3% to 13%, which is much less than those obtained when using also future context.

7. Analysis

This section illustrates the regularities that our models learned, and discuss the strengths and limitations of prediction from context alone.

For most dialog markers, the most predictive features were the pitch disalignment features. For most windows, these features had correlations of 0.20 or higher with high pitch, high volume, and high harmonicity. This is likely because peak disalignment often marks times of shared laughter, questions, and other high-engagement dialog acts [35], and these generally call for enthusiastic dialog markers. There was also a tendency to matching: more specifically, when the immediate past context exhibits higher volume or pitch, the prosody of the dialog marker often does too, for example when acknowledging new information. Some specific dialog markers had additional unique tendencies. For example, for the word now, high pitch correlated with high pitch by the same speaker over the next few sections, likely due to its forward-looking role, as in introducing new subtopics. Most of the strong correlations were with contextual behavior by the person who produced the dialog marker, but there were also interlocutor effects. For example, the word okay tended to be lower in pitch when in the context the interlocutor’s cepstral flux was low, likely due to the use of lengthening and reduction marking a low density of new information or seeking only weak feedback.

For insight on why the model sometimes performed well and sometimes poorly, we start by considering Table 3. We note relatively high predictability for uh, and huh, likely because they usually have no independent prosody or meaning beyond their roles in the local context. We see low predictability for mean features of really and for both mean and max features of right, which are sometimes dialog markers, but sometimes just adverbs and adjectives, in which roles they likely have different prosodic tendencies. The prosody of okay was also hard to predict, perhaps because it often is deployed to convey a specific meaning or function, rather than just fitting passively in the context.

To further understand where our model succeeded and failed, we examined its performance on specific tokens: for each dialog marker type, the 5 for which the predictions were least accurate, and the 5 for which they were most accurate. This was done subjectively, relying on our perceptions and qualitative inductive methods.

Factors that were common when the model failed included: i) background noise in the audio segment. (Our feature computations were not robust to noise.) ii) long monologues (a dialog activity type uncommon in Switchboard, and likely rare in the training data). iii) one speaker with an unusual accent or perhaps a speech impediment, iv) incorrect annotations, for example, where the label was um, but the sound was more like hmm. (Our model for um, of course, had not been trained to predict the prosody of hmm tokens.) v) sequences of dialog markers, such as well, yeah and oh, okay. (The prosody of markers in sequence is apparently different from those in isolation, the more common case in the training data.) vi) okay at the end of conversation, where it was short and breathy as part of the closing and vii) huh when produced as a repair question or strong exclamation.

Cases where the model’s predictions were most accurate included i) typical backchannel uses of yeah, ii) times the speaker and interlocutor shared happy or excited agreement, for example, You’re pretty Texan, yes …[interlocutor laughter], and iii) sympathetic produc-
Table 3: Results per Dialog Marker: Percent reduction in root mean squared error for predicting mean (respectively maximum) feature values using linear regression.

<table>
<thead>
<tr>
<th></th>
<th>predicting mean features</th>
<th>predicting maximum features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>le</td>
<td>cf</td>
</tr>
<tr>
<td>huh</td>
<td>29</td>
<td>43</td>
</tr>
<tr>
<td>now</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>oh</td>
<td>17</td>
<td>27</td>
</tr>
<tr>
<td>okay</td>
<td>18</td>
<td>30</td>
</tr>
<tr>
<td>really</td>
<td>46</td>
<td>16</td>
</tr>
<tr>
<td>right</td>
<td>37</td>
<td>2</td>
</tr>
<tr>
<td>uh</td>
<td>43</td>
<td>21</td>
</tr>
<tr>
<td>uh-huh</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>um</td>
<td>31</td>
<td>22</td>
</tr>
<tr>
<td>well</td>
<td>25</td>
<td>24</td>
</tr>
<tr>
<td>yeah</td>
<td>38</td>
<td>25</td>
</tr>
<tr>
<td>yes</td>
<td>30</td>
<td>19</td>
</tr>
<tr>
<td>Average</td>
<td>29</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 4: Results for predictions using only past context.

<table>
<thead>
<tr>
<th></th>
<th>predicting mean features</th>
<th>predicting maximum features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>le</td>
<td>cf</td>
</tr>
<tr>
<td></td>
<td>Model RMSE</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Reduction, %</td>
<td>6.0</td>
</tr>
</tbody>
</table>

8. Discussion and Future Work

We have found evidence that dialog markers’ prosody can indeed be predicted directly from the prosody of the context, to some extent. This is true even with a very limited feature set and a very simple model.

Much better performance should be obtainable using better models and more context features, including not only more prosodic features, but also lexical information. Future work should also attempt more detailed predictions: not just of a token’s averages, but also or contour parameters or even frame-by-frame values.

An important open question is the extent to which the prosody adjustments recommended by a context-sensitive model have actual value in dialog. Various previous research suggests that improved responsiveness can improve perceived naturalness, responsiveness, and ultimately rapport, engagement, and user satisfaction [25, 36, 37, 38, 39], but human subjects experiments are needed to establish whether this is also true here.

This will pave the way to more responsive dialog systems, for example spoken language chatbots. Optimal exploitation of context-based prosody predictions may, however, require advances in speech synthesis, to support generation of tokens that exhibit fully appropriate prosody.

While simple context-based control of dialog marker prosody may be adequate for chatbots, where the aim is to keep the dialog flowing, use in task-oriented systems will bring further challenges. We would likely need an additional module trained to judge the similarity of the local dialog context to the context in the training data. We would also need methods to combine these context-based predictions with other factors that must affect the prosody, such as the current dialog state and the communicative intent of the system [40].

More generally, future work should explore the possibility of predicting other aspects such as the prosody of full utterances in dialog based on local context. Such models could help not only improve dialog systems, but might also be used to help autistic people and language learners master the typical patterns of responsiveness in dialog, thereby helping improve their communication skills.
9. References


[34] ——, Prosodic Patterns in English Conversation. Cambridge University Press, 2019.


