Lexico-Prosodic Anomalies in Dialog

Nigel G. Ward, Alejandro Vega, David G. Novick

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

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

Abstract

In dialog, most words fit nicely with their prosodic context. This enables the prediction of which words are likely to come next, given the recent prosodic context, an ability which is of practical utility. However anomalies exist, cases where the word that comes next seems to be a mismatch for its prosodic context. Examination of 60 such lexico-prosodic anomalies in the Switchboard telephone dialog corpus revealed some patterns: anomalies tend to occur with, or perhaps constitute, bids for dominance, expressions of emotion, idiosyncratic speech patterns, and prosodic amalgams.

Index Terms: predictions, unpredictable, unlikely, language model, dominance, prosody

1. Prosodic Anomalies

The study of various kinds of anomaly in speech has a long history [1]. Speech errors and slips of the tongue do not occur randomly, but fall into specific patterns. Discovering and modeling these patterns has been a primary source of insight into the cognitive structures and mechanisms of speech production.

Most work on anomalies has focused on lexical and phonetic mishaps; studies of prosodic anomalies have apparently been limited to errors of lexical stress and the way in which prosody interacts with false starts and recoveries [2]. Two key problems hinder the study of prosodic anomalies. First, purely prosodic errors are vanishingly rare. Generally a prosodic pattern is strange only if it fails to align with the pragmatic and semantic intent or with the syntactic and lexical properties of the carrier phrase. Second, the field lacks clear criteria for whether a suspect prosodic pattern is actually an error. The identification of lexical errors is relatively easy: one writes down what one hears, and if the result is sequence of letters that fails to be a word, or a sequence of words that fails to be grammatical and meaningful, then it can be considered an error. Prosody, however, lacking agreed-upon symbolic representations and clear norms, let alone formal rules, is not amenable to this method.

This paper reports an initial exploration of prosodic anomalies. We solve the problem of identifying them by using a model of the mutual appropriateness of words and prosodic contexts. This enables us to find cases which, although not errors in the sense of being rule violations, are anomalies in the sense of being highly unlikely according to the model.

The paper is structured as follows. Section 2 explains how prosodic information can be of value for the problem of predicting upcoming words in dialog, and Section 3 outlines a model which does so. Section 4 explains our notion of lexico-prosodic anomaly and Section 5 explains the methods of analysis. Section 6 presents what was found with the Switchboard corpus, including the connection between these anomalies and violations of social norms, and Section 7 discusses the significance and potential utility of these findings.

2. Prosody-Based Word Prediction

In human interaction, the ability to predict the actions of an interlocutor at the micro-level, moment-by-moment, has been identified as a central issue in coordination, and better predictions correlate with more empathy and success in interactions [3, 4, 5, 6]. In dialog, prosody plays a major role in enabling dialog participants to predict each others' actions, for turn-taking, obviously, but also in other respects. One aspect of this is the role of prosody in helping people (and machines) accomplish the feat of word recognition: prosodic context provides expectations that help the hearer recognize each upcoming word quickly and accurately.

The power of prosody to aid word prediction relates to the engineering problem of language modeling; that is, the problem of predicting the speaker's next word, given the previous words and other prior context. Having good language models is important, not least because every speech recognizer relies on one to provide probability estimates for the word hypotheses it considers. Mainstream language models use only the *lexical* context, and it is probably not coincidental that, despite substantial recent progress, speech recognizer performance is still weak for spontaneous speech in general and dialog in particular. However experimental evidence suggests that the potential informativeness of signal and interlocutor-track information exceeds that of just more lexical context [7], and that it is often the prosody that provides the information that enables better predictions.

While prosodic features tied to lexical and syntactic context can help predict the upcoming word in broadcast speech [8, 9, 10, 11], our interest here is in dialog, where the effects of lexical and syntactic factors on the prosody are often swamped by the effects of cognitive and interpersonal factors, such as delays while thinking of the right word and the management of who speaks when.

For example, a speaker who takes up the turn immediately after the other has ended may not be ready to produce a fluent utterance, so his or her speech may start with a disfluent region. Such regions may be recognized by reduced pitch range, slow speaking rate, and low volume, among other features. In such regions high-content words are less likely, however, as the speaker continues formulating, content words are likely to appear before too long. The listener, perceiving these features, knows what to expect; indeed, the production of disfluency-marking prosody can serve as a communicative strategy for informing the listener that he or she is not to take the floor, nor to pay much attention to the words produced while in this state [12].

3. A Predictive Model

The model used here was originally built for another purpose, to improve speech recognition using prosodic information. Ob-

serving that certain words become more or less common in various prosodic contexts, we set out to use prosodic information in the context up to time t to predict which word will occur starting at time t [13, 14, 15]. We tried various prosodic features, mostly those implicated in the literature in the expression of cognitive states, communicative functions, or both.

We trained various models based on these features, using about 650,000 words of dialog from the Switchboard corpus, a collection of telephone conversation of spontaneous topics between unacquainted adults, as transcribed by ISIP [16, 17]. In many cases we found interesting tendencies, some easier to understand than others. For example, after low-volume regions the likelihood increases for words such as *true, definitely, might, tend, mostly* and other terms relating to belief. After regions of increased speaking rate, place names, numbers, and other content words become more likely.

To date the best models use eight features. Four are local prosodic features computed over small time windows immediately preceding the word to predict,

- speaking rate (previous word's shortening/lengthening)
- volume (over the previous 50 milliseconds)
- pitch height (over the previous 150 milliseconds)
- pitch range (over the previous 225 milliseconds)

and the other four are the times elapsed since various prosodically-marked dialog events,

- time into utterance
- time since other's most recent utterance end
- time since own most recent low pitch region
- time since other's most recent low pitch region

The output of the model is, for each word, a ratio indicating how likely that word is in that context: greater than 1 if more likely there than usual, and less than 1 if less likely. The likelihoods are computed using the strategy and methods described in [15]. The model used here incorporates two recent improvements: the speaking-rate estimates are based on the word durations rather than an rough acoustic estimate, and the local prosodic models make no contribution when the word to be predicted is the first one of an utterance (coming after at least 1.2 seconds of silence). The weights of the models were uniformly 0.3; optimization was not done. This model was among those that performed best in making accurate predictions. Judged using the usual metric, perplexity, the incorporation of prosodic information gave a 4.6% reduction relative to a trigram baseline. The model performed well on average, and 65% of the words in the test set had their estimates improved by the model.

An example of a success is the words *sounds* in *that sounds nice*, produced as a quiet comment in response to talk about a ski resort. The contributions of each feature to the overall likelihood ratio were:

1.00	speaking rate on *that* was faster than average
1.26	volume was quiet
1.12	average pitch was relatively high
1.00	not enough pitch points to reliably estimate the range
1.18	located about 600 milliseconds into an utterance
1.27	located about 900 milliseconds after the interlocutor ended his utterance
1.00	located about 5 seconds since speaker's own last low pitch region
1.15	located more than 9 seconds since the interlocutor's own last low detected pitch region
0.98	from the normalization

Normalization had a negative impact here because other words in the vocabulary also received overall likelihood boosts; specifically, the weighted average of all such boosts was 1.02, where the average was weighted by the trigram probabilities in this context.

4. Lexico-Prosodic Anomalies

Given this model, *lexico-prosodic anomalies* are cases where a word appears in a prosodic context where it is unlikely according to the model. Although the model was not designed for finding anomalies, it does have properties that make it suitable for this purpose.

First, the prosodic features used in the model are computed directly from the acoustic signal and the word-aligned transcript. (While the word labels were used to help compute three features — time since utterance start, time since other's utterance end, and speaking rate — these could have instead been computed directly from the acoustic signal.) Thus there is no need to use a corpus with hand-labeled prosodic features, which makes it possible to use a model tuned on a large dataset. This in turn reduces the danger of flagging something as an anomaly just because it happened that nothing like it occurred in some small training set.

Second, the model has been ruthlessly trained to achieve good performance. As a side effect, the parameters and weights are such that the model makes strong predictions only when there is strong evidence that the word in question in fact appears in the given context significantly more (or less) frequently in that context than on average.

Third, the model incorporates multiple features, making its predictions more robust. This is important because values for individual prosodic features are noisy. For example, a stressed syllable may lack the expected pitch peak if the stress is instead realized with energy or duration features. The model combines the contributions of all features into a single number, a probability estimate. When this estimate is very low for a word that in fact occurs, that word is generally a mismatch for the context on several prosodic features, and thus probably a true anomaly.

Finally, the model combines the contributions of the features by multiplication. If for example the duration of the previous word indicates that the word *people* is 1.2 times more likely here than usual, and in addition the pitch range over the previous 225 ms indicates that *people* is 1.2 times more likely here, then the two models combined indicate that it is 1.44 times more likely overall. This simplistic method assumes, incorrectly, that the various sources of information are independent, which sometimes hurts the predictive power but makes it easy to analyze why the model considers any specific case to be an anomaly.

5. Method

To explore the nature of these lexico-prosodic anomalies, we examined the 60 most anomalous tokens in a test set of 29,966 words from the Switchboard corpus. These 60 words had the lowest probability estimates according to the model described in Section 4. Given the specific prosodic context in which they occurred, these tokens were considered by the model to be quite unlikely (3.5–9.9 times less likely in that context than in Switchboard overall); and yet in fact they occurred.

In examining the anomalies we sought patterns, especially patterns that might point to weaknesses of our model that could then be fixed to improve its predictive power. Specifically, we examined each anomaly in three ways. First we examined the contributions of each of the eight features. In some cases we ex-

amined the model to determine why it indicated that a word was uncommon in a specific context; that is, we sometimes looked to see in what contexts the word more typically occurred in the training data. Second we looked at the lexical context in the dialog. Third, we listened to the audio.

We noted the factors that seemed to be behind each anomaly and the patterns that began to emerge. We considered possibilities of every type we could think of: problems with our model, mismatches between training data and test data, problems in the way our code computed the prosodic features or the likelihoods, peculiarities of the context, unmodeled interactions with other aspects of language (lexical stress, syntax, collocations, dialect), dialog acts, speaker cognitive states, interpersonal dynamics, and so on. Thus the method was inductive and largely qualitative.

6. Causes of Anomalies

Our analysis found about ten principal patterns. Some anomalies were involved in more than one of these patterns, thus there is some double counting of the 60.

Dominance was involved in twelve of the anomalies. Five of the anomalies were aggressive turn grabs (or turn-grab attempts), for example *yeah but it's close to Philadelphia*, in which *but* is unusual because it comes after a very fast *yeah* that is also loud and high in pitch, and because it happens less than 200ms into the utterance. Five anomalies involved speakers effectively performing monologs, which led to anomalies on words following long, dramatic, fillerless pauses, and on words where the speaker was agreeing with what he himself had just said, for example at the first *yeah* in *when you combine with the with the mismanagement that a lot of American companies have had and yeah I think yeah*. Two anomalies seemed to reflect idiosyncratic lexical uses, which also, to our ears, expressed dominance: *um* as part of a floor grab in *well, um*, and *huh* as question particle in *you did it recently too, huh?*,

Mock quotations such as *but people think, oh we're in the nineties, we're beyond all that*, were involved in twelve anomalies. In this example for *oh* there was a mismatch with the prosodic context: the model did not consider *oh* likely after a region of fast, loud speech with high average pitch, given that in Switchboard *oh* is predominantly a true discourse marker. These were not literal quotations, but rather words illustrating what someone was thinking or saying, including the past thoughts of the speaker himself or herself.

Departures from the norms of dialog were involved in ten anomalies, in various ways. Most reflected a shift of a speaker to a monolog style, and of these most seemed to also be due to one speaker taking a dominant role, as noted above. There was also one case occurring during a return from an aside to a third party, and two cases of a speaker apparently indulging in a dreamy reminiscence.

Emotion was expressed in or near nine anomalies. For example, this was seen were a speaker responded to news about a pitcher's injury with *oh jeez*, where *jeez* typically occurs after words that are fast, low volume, and have a narrow pitch range. Another example was the word *remember* in *because children [500ms-inbreath] remember the the traumas*, where *remember* generally does not occur after slow, quiet, pitch-less regions. There were also five anomalies in mock quotations that conveyed emotion, as at *oh* in *a little more self-confidence built up, oh yes I can do this*.

Words with multiple senses or uses accounted for eight anomalies. For example, the word *right* was anomalous in both *I'm having dinner right now* and *fit right in with the PVC*, as in Switchboard *right* is almost always a discourse marker, and thus the model judged it unlikely to occur in the middle of a fluent segement. Another example was the word *hours* in the phrase *office hours*. In Switchboard the word *hours* typically occurs after a stressed syllable, as in *one hour*. The model picked up this regularity, and thus this use of *hours* was flagged as anomalous.

Corpus bugs caused eight anomalies. There were a few cases of cross-track bleeding, which interfered with the local prosody computations. There were also five mislabelings, including mislabelings of the purportedly anomalous word (for example, *uh-oh* mislabeled *huh oh*), mislabelings of the previous word (which interfered with the speaking-rate computation), and misplacement of the word-onset timepoint (which interfered with the windows over which the local prosodic features were computed).

Completing the other speaker's phrase was a factor in six anomalies, as in *then be gone* coming 170ms after the other speaker had said *the sticky buns of course would last three or four days but, and that'd be it*.

Sarcasm was involved in five anomalies, including three in mock quotations.

Prosodic amalgams occurred in four of the anomalies. Although a person in conversation may be thinking of many things simultaneously, usually he or she selects just one of them to express, so that each utterance is internally consistent and relates to a single thought. When this is not possible, for example when a speaker begins without having resolved what he or she wants to say, this internal conflict is usually well signposted prosodically. However there were anomalies that seemed to reflect a radical, swift "change of direction" without overt prosodic signposting. For example, this was seen at *Cowboys* in *um that's um Cowboys are going to have a problem*. In this dialog, the attempt to find common ground on the topic of football had faltered, and the speaker appeared to be trying to shift to a new topic baseball, but then, we suspect, suddenly remembered a recent news tidbit about a football player trade, and blurted it out starting with the word *Cowboys*. The transition from a slow speaking rate with small inter-word pauses to the word *Cowboys* is highly unusual in Switchboard, as it probably is for high-content words in general.

Other factors were involved in a few anomalies. Two instances seemed to acknowledge acceptance of the dominant role of the other speaker. One was apparently a simple disfluency. One may have been a case of self-monitoring, where the speaker was unsure whether it was appropriate to disclose some information. And, one is a complete mystery at this point; the prosodic features did not significantly affect its estimate, but the normalization did, meaning that some other word or words were much more likely according to the model, but with our current tools we cannot determine which or why.

Finally, orthogonally to the patterns noted above, many anomalies occurred **late in the dialog**. For example, 42% occurred between 200 and 300 seconds into the dialog, while only 24% of the tokens overall were in this range (significant $p < 0.005$, χ^2). After some time into the dialogs some of the speakers were opening up to each other, while in other cases the speakers were losing interest. More generally, with time it seems that the conversants began to stray from cultural norms and behave more idiosyncratically.

7. Conclusions and Directions

As hoped, the lexico-prosodic pairings flagged by the model as anomalous generally do seem to be truly anomalous, in that they appear to be atypical of the telephone-smalltalk genre that dominates Switchboard. For example, it found direct contradictions among largely agreeable exchanges; competitive turn openings among largely relaxed dialogs, moments of distraction among largely attentive listening, domineering one-sided dialogs among largely balanced interactions, displays of emotion among mostly calm exchanges, completions of the other person's sentences among mostly arms-length turn-taking, and mislabelings in a generally high quality corpus. Such correlations between model-detected anomalies and truly anomalous dialog events provide some validation for the model. This finding may also have practical significance. Recently there has been interest in automatically detecting dominance relations from dialog data and in finding regions of high participant involvement ("hot spots") [18, 19]. Locating lexico-prosodic anomalies may be useful for such purposes.

From a psycholinguistic perspective, the discovery of amalgams, where two different prosodic contours abut without overt disfluency markers, is interesting. Examination of these phenomena may provide insight into the mental representation of prosodic intentions and the mental process of monitoring prosodic output.

As a side-effect of this study, we identified some possible ways to improve our predictive model. Examination of the anomalies indicated that the model's performance is indeed weakened by the aspects of prosody that are not included, notably turn-hold signals (especially before long pauses) and lexical stress. The pattern of the prosodic features involved in some anomalies further suggests that the independence assumption sometimes hurts performance. For example pitch range, pitch height, and volume all correlate, meaning that a word that is unlikely in the eyes of one of these features is often unlikely for all, meaning that its probability estimate is penalized three times over, again hurting performance.

Ultimately what we want is a deeper model, that goes beyond the partially redundant surface features to represent the actual information that prosody is providing at any given point in a dialog. This would be useful for many purposes, but is a long-term challenge for the field.

The model we used to relate prosody and lexical choice was a predictive one, and so it could only identify anomalies in which there was a mismatch between a word and its left context. It would be interesting to extend this work to explore mismatches between a word and the prosodic realization of that word itself, and to explore mismatches between the word and the following prosody. Looking further ahead, we could flip the predictive model around, using various factors to predict the prosody at each point, and then see what could be learned by from mismatches between predicted prosody and actual prosody in the corpus.

Although our model and our method have limitations, we have developed a model-based technique for identifying lexico-prosodic anomalies, shown that prosody in dialog is largely predictable (with the significant departures from standard patterns generally not occuring without reason), and pioneered a new approach that may give insights into various aspects of prosody in actual use.

8. Acknowledgments

This work was supported in part by NSF Award IIS-0914868.

9. References

[1] V. A. Fromkin, "The non-anomalous nature of anomalous utterances," *Language*, vol. 47, pp. 27–52, 1971.

[2] J. M. Brenier and L. A. Michaelis, "Optimization via syntactic amalgam: Syntax-prosody mismatch and copula doubling," *Corpus Linguistics and Linguistic Theory*, vol. 1, pp. 45–88, 2005.

[3] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. van der Werf, and L.-P. Morency, "Virtual rapport," in *6th International Conference on Intelligent Virtual Agents*, pp. 14–27, 2006.

[4] L. W. Barsalou, C. Breazeal, and L. B. Smith, "Cognition as coordinated non-cognition," *Cognitive Processing*, vol. 8, pp. 79–91, 2007.

[5] E. Jahr and S. Eldevik, "Response variability and turn taking in cooperative play," *Journal of Speech and Language Pathology*, vol. 2, pp. 190–194, 2007.

[6] J. Streek and J. S. Jordan, "Projection and anticipation: The forward-looking nature of embodied communication," *Discourse Processes*, vol. 46, pp. 93–102, 2009.

[7] N. G. Ward and B. H. Walker, "Estimating the potential of signal and interlocutor-track information for language modeling," in *Interspeech*, pp. 160–163, 2009.

[8] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Mathematical Foundations of Speech and Language Processing, IMA Volumes in Mathematics and Its Applications, Vol. 138*, pp. 105–114, Springer-Verlag, 2004.

[9] K. Chen, M. Hasegawa-Johnson, and J. Cole, "A factored language model for prosody-dependent speech recognition," in *Robust Speech Recognition and Understanding* (M. Grimm and K. Kroschel, eds.), pp. 319–332, I-Tech, 2007.

[10] S. Huang and S. Renals, "Modeling prosodic features in language models for meetings," in *Machine Learning for Multimodal Interaction IV (LNCS 4892)* (A. Popescu-Belis, S. Renals, and H. Bourlard, eds.), pp. 191–202, Springer, 2007.

[11] S. Ananthakrishnan and S. Narayanan, "Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework," in *ICASSP*, pp. 873–876, 2007.

[12] R. J. Lickley and E. G. Bard, "On not recognizing disfluencies in dialogue," in *Interspeech*, pp. 1876–1879, 1996.

[13] N. G. Ward and A. Vega, "Modeling the effects on time-into-utterance on word probabilities," in *Interspeech*, pp. 1606–1609, 2008.

[14] N. G. Ward and A. Vega, "Towards the use of inferred cognitive states in language modeling," in *11th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 323–326, 2009.

[15] N. G. Ward and A. Vega, "Using non-lexical context to improve a language model for dialog," *Speech Communication*, 2010. submitted.

[16] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of ICASSP*, pp. 517–520, 1992.

[17] ISIP, "Manually corrected Switchboard word alignments." Mississippi State University. Retrieved 2007 from http://www.ece.msstate.edu/research/isip/projects/switchboard/, 2003.

[18] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations from non-verbal activity cues," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 501–517, 2009.

[19] B. Wrede and E. Shriberg, "Spotting 'hot spots' in meetings: Human judgments and prosodic cues," in *Eurospeech*, pp. 2805–2808, 2003.