Response Types and The Prosody of Declaratives

Catherine Lai

Department of Linguistics, University of Pennsylvania, Philadelphia, USA
laic@ling.upenn.edu

Abstract

This paper presents a production study of declarative responses in varying dialogue contexts. The goal was to determine what sort of dialogue conditions produce systematic variations in prosody, and what these variations mean for discourse interpretation. After controlling for information structural factors, we found that distinct prosodic forms were predictably and consistently elicited by varying the response type of the utterance. In particular, we found that indirect agreements/contradictions were produced with a distinct intonational form compared to to direct responses. We quantify the prosodic separability of these response types via classification experiments, comparing the usefulness of both aggregate features, e.g. mean and variance, and features derived from function decomposition techniques. We find that the latter approach allows us for a more succinct description of category differences in terms of tilt and convexity.

Index Terms: Prosody, Production, Pragmatics, Dialogue acts, Information structure, F0 modelling.

1. Introduction

This paper investigates the relationship between prosody and dialogue structure. That is, what sort of dialogue conditions produce the systematic variations in prosody, and what do these variations mean for discourse interpretation. Past studies have found prosodic features to be useful for recognizing discourse categories like dialogue acts [1, 2]. However, it unclear how generalizable these connections are. For example, [1] find that affirmative backchannels have rising pitch in game oriented dialogues, while [3] do not find the same in meeting data. Moreover, while this sort of approach is informative for recognition, it avoids the crucial question: what are those features doing there in the first place? This is something we need to know if we want to tackle, for example, the problem of synthesizing expressive, conversational speech.

Corpus based studies like the ones cited above usually describe distributions of features over utterance level categories like dialogue acts. On the other hand, theoretical analyses often put forward direct mappings between intonational forms and meaning. For example, analyses in [4, 5] try to relate pitch accents shapes and boundary tones to specific informational structural (IS) categories like topic and focus, i.e. units below the dialogue act level. Although experimental evidence suggests that these mappings are too tight, IS does does seem to play a consistent role in determining the prosodic form of an utterance. For example, based on experimental data [6] argues that the topic/focus difference can be characterized in terms of relative prominence instead of accent/boundary shape.

With this in mind, we would like to know how both higher level discourse structure and IS requirements change the expectations of what intonational forms are available. Impressionistically, declarative responses like (1b) and (2b) are produced with different contours even though they have identical propositional content and perform the same speech/dialogue act (assertion/statement/inform).

(1) a. So, Emily brought a meringue.
    b. Right. Emily did bring a meringue

(2) a. Nobody brought a dessert.
    b. Emily did bring a meringue

(3) a. Emily brought a meringue.
    b. What? Emily did bring a meringue?

The intuition is that the *direct agreement* in (1b) is produced with falling pitch from 'did', while the *indirect contradiction* response (2b) naturally has an extra fall-rise accent on 'meringue'. Furthermore, using the intonational contour of (1b) sounds infelicitous in the context of (2). Both of these differ from the declarative question (check move) in (3b), which is intuitively produced with a rising accent. So, it seems that these sorts of response dimensions may be more consistent predictors of meaningful prosodic variation than the usual type of dialogue act. However, to support this idea we need to how robust these intuitions are. Furthermore, we would like to know whether the intonational form is affected by whether response is an agreement or contradiction.

To investigate these issues, this paper presents a production experiment which examines the relationship between response types and prosodic form. Section 3 presents the results given the experimental setup described in Section 2. The results show that we can distinguish direct declarative responses from indirect responses, and similarly from declarative questions, based on the F0 contour shape at the end of the utterance. Section 4 presents classification experiments with the goal of quantifying how separable the prosodic features associated with these classes are. We compare classifiers based on different sets of aggregate features, e.g. F0 mean and variance, as well as features drawn from function decomposition techniques. Using the latter approach, we see that the intonational differences between categories fall out more clearly when viewed via shape features like tilt and convexity. Section 5 discusses some implications of this approach. Section 6 concludes.

2. Data and Method

The production experiment consisted of two parts, both recorded in pairs: two scripted dialogues (long ≈ 30 turns each) and statement/response pairs (short). The target responses in the latter task consisted of two declaratives with verum focus (E: 'Emily did bring a meringue', M: 'Marianne did meet with Lenny') and two with broad focus (W: 'William ran away', Y: 'Mary remembered your birthday'). Four distractor sentences were also included. The motivation for using these constructions was to examine the relationship between dialogue structure and prosody while keeping the IS fixed. In verum focus

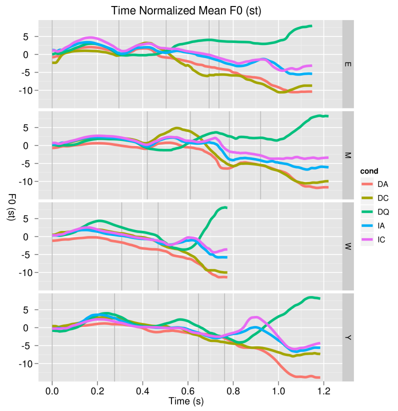

Figure 1: *Mean F0 contours (Time normalized per word). The bars indicate average word start time.*

sentences, IS focus is the proposition's polarity. This draws main sentence stress from the default utterance end position to the inserted 'did'. So, unlike the broad focus cases we consider, post 'did' accents will mark units in the IS ground (or theme) rather than the IS focus. The pairs were designed to cover direct (D) and indirect (I) agreements (A) and contradiction (C) type responses, as well as declarative questions (DQ). This resulted in 20 ({E,M,W,Y}x({D,I}x{A,C}+{DQ}) target stimuli of similar form to (1)-(3), read twice by each participant.

The scripted dialogues involved two scenes where the participants talked about a past event. The scenarios were set up to elicit direct and indirect agreements and contradictions. Four turns reproduced conditions from the short context recordings (EIC, MDA, WIA, YDC). Again, each speaker recorded each part twice. However on the second recording participants were asked specifically to try to sound more involved/engaged in the scene. All recordings were recorded in a sound attenuated booth. Eight pairs of speakers of Standard American English participated (7 males, 9 females).

Timing data was initially obtained by using using the Penn Phonetics Lab forced aligner [7], after which word and turn boundaries were manually corrected. F0 and Intensity features were extracted via Praat. F0 contours were also manually corrected trimmed and smoothed via Xu's ProsodyPro Praat script [8]. These values were normalized to a semitone scale relative to each speaker's median F0. Intensity and duration measurements were converted to z-scores by speaker and word respectively. The following aggregate features over F0 and intensity were calculated over each word: mean, standard deviation, slope, jitter, maximum, minimum, absolute range, absolute and relative times of maximum value (minimum similarly).

3. Results

3.1. Short Contexts

Figure 1 shows mean time normalized F0 contours for the short context target sentences. We see immediately see that most of the variation between response types happens at the end of the utterance. As expected direct responses generally fall through the final word while the declarative questions rise. The indirect responses show a fall-rise shaped accent on the final word for sentences E, W, and Y, while for sentence M, the fall-rise is spans three words from 'meet' to the end of the utterance. This is expected since the verb is contrastive in that context ('I think Marianne is conspiring with Lenny').

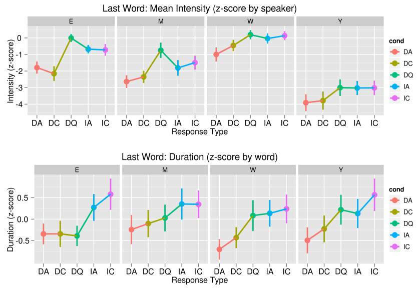

Figure 2: *Last word intensity and duration by response type*

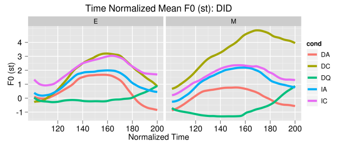

Figure 3: *Time normalized mean F0 (st) contour for the verum focus particle 'did'.*

Pairwise t-tests over last word features showed significant differences between direct, indirect and question responses in terms of the F0 mean, min, max, and relative min and max times over the last word ($p < 0.01$, Bonferroni corrected). The declarative questions were also significantly different from all other classes in terms of slope. The different response types induce less variation in intensity. However, it does appear that the indirect and declarative question productions have a higher mean intensity for the utterance final words (t-test, $p < 0.01$). In the same vein, we see relatively longer final word duration for question and indirect responses (Figure 2).

We do not see similar differences at the contour tail on the agreement/contradiction dimension. Pairwise t-tests did not show any significant differences comparing the last word features of the indirect agreements and contradictions. However, direct agreements and contradictions did show a significant difference in slope ($p < 0.05$). This reflects the fact that for broad focus contradictions some speakers made use of Sag and Liberman's 'contradiction contour' [9]. This did not appear to be an option for the verum focus contradiction.

We see a more consistent difference between agreements and contradictions when we look at the F0 contour on the locus of verum focus: 'did' (Figure 3). In fact, direct contradictions had significantly higher F0 mean, standard deviation, slope, jitter, maximum value and relative peak time than their agreement counterparts. No significant differences with respect to intensity and duration features were found, so pitch seems to be main carrier of this difference. Note, even though polarity is clearly contentious in this context, the accent appears as a (delayed) peak contra Steedman's [4] claim that H* pitch accents mark uncontentious informational units. Overall, contradictions have the same shape as agreements but employ a more emphatic, effortful gesture.

3.2. Long Contexts

The mean contours for the longer scripted productions are as expected given the response type. In particular, we see somewhat more articulated fall-rise shapes on the indirect responses (E, W). We also see more use of the contradiction contour for the broad focus contradiction (YDC). The use of this contour varied between and within speakers. Figure 4 shows this YDC contour for two speakers: mm1 used the contradiction contour

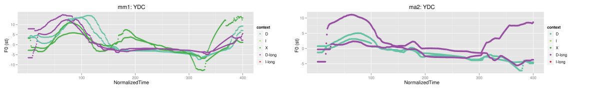

Figure 4: *Use of the contradiction contour: ‘Mary remembered your birthday’*

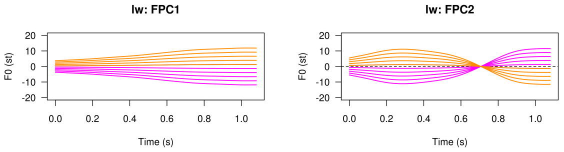

Figure 5: *The first two FPCA harmonics (fda.lw). Orange/Magenta = coefficients +/- 1,3,..,9 .*

consistently while ma2 used it only once in the longer context recording. Overall, the increased use of this more exaggerated contour is inline with the idea that speakers were more involved in the long dialogues.

3.3. Functional Features

Although we can capture the differences in response type in terms of aggregate features, the differences seem to more about shape of the last pitch accent: D, I, and Q categories map to fall, fall-rise and rise accents respectively. Similarly, we see that the former two assertive moves have concave accent shapes, while the latter check move is convex. However, adherence to these canonical shapes is clearly not strict. As we have seen, broad focus direct contradictions can have rises through the final word. Similarly 46% of indirect responses had a negative slope in the last 100ms.

We employ Functional Principal Components Analysis (FPCA) to better capture these shape variations. Like regular PCA, we use this technique to describe our contour as a linear combination the of principal components/harmonics. In this case these harmonics are functions which represent the dominant modes of variation in the data. We follow [10] in deriving the FPCA scores (i.e. harmonic coefficients). The data was initial fit using B-splines and FPCA was performed on the short context productions using the R package `fda`. We looked at two time domains: the last word (fda.last) and the segment from the last actually prominent word at or after the expected main stress position (fda.lw). So, for example, we take *‘meet with Lenny’* for sentence M indirect responses, but only *‘meringue’* for sentence E. This was done to capture the fact that the accent and tail may span multiple words.

Figure 5 shows the first two harmonics of the FPCA (fda.lw). These two modes describe the tilt and convexity of a contour and accounts for 80% and 15% of the variation in the data respectively. The projection onto these two dimensions (Figure 6) matches our impressionistic analysis. In particular, the second harmonic scores reflect the prevalence of the fall-rise shape at the end of the indirect responses, scooping rises for the question type responses. The PCA projection in Figure 5 only really distinguishes the response types on the first component dimension which is dominated by mean F0. That is, the PCA based on the F0 aggregate features doesn’t pick up on the these differences in convexity.

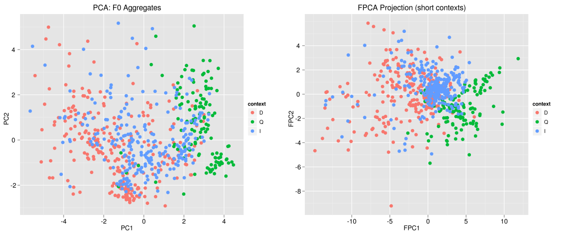

Figure 6: *PCA (F0 aggregates) v FPCA projection (fda.lw)*

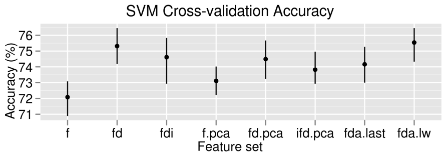

Figure 7: *10 fold cross-validation accuracy (short context data)*

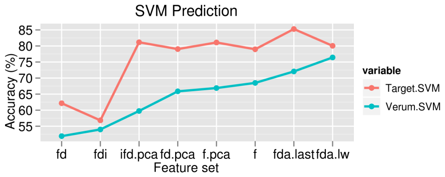

Figure 8: *Prediction Accuracy on long context data*

4. Classification Experiments

In this section we examine how separable the response types are with respect to different prosodic features. To do this we compare the performance of classifiers trained on different feature sets from the short context set. Additionally, we examine the robustness of the classifiers with respect to unseen data from the long context productions. We compare classification into direct (D), indirect (I), and question (Q) response classes based on the F0 (f), intensity (i) and duration (d) features, PCA based transformations of those feature vectors (e.g. fdi.pca), as well as the scores from the first five harmonics derived from the FPCA decomposition (fda.last, fda.lw). Support Vector Machine (SVM) classifiers (with radial basis function kernel) were trained for each of the feature sets on the short context data using the R package `e1071` (LIBSVM) with hyperparameters optimized using a grid search.

Figure 7 shows mean 10 fold cross-validation accuracy scores and 95% confidence intervals over 100 randomized runs. The FPCA (fda.lw) classifier and the aggregate F0+duration (fd) classifiers performed the best at around 75% accuracy. The worst performance came from the F0 only classifier (f) at 72%. All classifiers performed significantly better than the 40% baseline of classifying all utterances as indirect responses. So, the response types are distinct but there is still seems to be considerable overlap between the classes.

Figure 8 shows how well the short context trained classifiers predict the response type of utterances from the long context recordings. The graphs shows the prediction accuracy for the E, M, W sentences (*Target-SVM*) which appeared in both long and short contexts (we leave the Y case for later), as well as 6 other direct and indirect responses involving verum focus, unseen with respect to the training data (*Verum-SVM*). The FPCA based classifiers perform the best for both these groups. Accuracy with respect to the Verum group is around the same level as the cross-validation results for the short context data (72-76%), while the Target group scores several points higher (e.g. 85% for fda.last). While the FPCA based classifiers perform better with the new data than those based on aggregates, it is worth noting that the F0 only classifier (f) is not that far behind. This

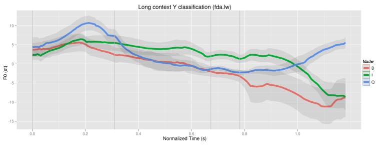

Figure 9: *Mean contours for predicted classes for sentence Y in the long context recordings (fda.lw).*

suggests that F0 characteristics are generally more robust indicators of response types than duration and intensity.

Finally, as noted previously, sentence Y was produced with the contradiction contour in a number of cases, especially in the long context dialogues. Figure 9 shows the mean contours and confidence intervals grouped by the fda.lw classifier. This classifier split the Y productions into three evenly sized groups (D=11, I=10, Q=11). We see that the Q labelled contours, indeed, have the shape of the contradiction contour. So, it seems there are more prosodic options for broad focus utterances, and specifically for contradictions, which this sort of quantitative approach can help tease out.

5. Discussion

The production study showed systematic prosodic differences based on whether a declarative was a direct or indirect response. These categories were broadly characterized by fall and fall-rise terminal contours respectively. If we were just to look at the relationship between prosody and their sentence type (declarative), illocutionary force (assertion), or the dialogue acts categories (statement, inform), we would not be able to see the patterns. This highlights the fact that to get coherent predictions about prosodic forms, we need to take the surrounding discourse structure into account.

Looking at this structure sheds light on why indirect responses are characterized by an IS unit having an unexpected level of prominence. In the verum examples, this is a second accent after the IS focus 'did'. For the broad focus utterances, we see a bigger gesture on the metrical main stress position. This extra prominence seems mark a contrastive element and evoke alternatives. For example, so (2b) provides a partial answer for the question 'Who brought what?'. Answering the bigger question is a *strategy* for refuting the claim that 'Nobody brought a dessert'. In fact, Büring [5] argues that fall-rise accents basically do this, i.e. signal a set of subquestions which are relevant for answering a question higher up in the discourse tree. The extra contrast gives a template for generating the strategy.

Büring, further argues that fall-rise accents mark IS units as being *contrastive topics*. However, experimental evidence suggests that such a strict mapping is not warranted. Like [6], our production data indicates that an actual fall-rise shape (as opposed to a fall), while common, is not necessary. By evoking a strategy, via the extra contrast, indirect responses implicitly leave the current question under discussion open. While this is congruent with the contribution of a terminal rise [11], the rise isn't necessary to get this implication of openness. In general, it seems that intonational forms like fall-rises reveal what the speaker thinks the current dialogue configuration is, but they can't actually force the dialogue to take on that configuration. That is, intonational units don't act like semantic operators. H* accents are no guarantee that an IS unit is agreed upon. Similarly, a fall-rise accent can appear on units other than contrastive topics. So, it seems that to really investigate the prosody-meaning map, we need a notion of dialogue move that reflects not just the illocutionary force of an utterance but also how it fits into the dialogue structure, i.e. response types.

6. Conclusion

The main goal of this paper was to show that we get a better understanding of the prosody-meaning map if we look at categories which more directly reflect dialogue structure and expectations. In the production study, we indeed found that prosody varied consistently by looking at responses dimensions like whether an utterance directly or indirectly addresses the question under discussion. While machine learning experiments show that we can separate out the different response types based on aggregate statistics like mean F0 to a fair degree, the differences in the prosody are more succinctly categorized by the overall contour shape from the last accent to the utterance end. We quantified these differences using functional data analysis techniques. This allowed us to quantify contour shape over a continuous space with minimal manual annotation effort: we only have to specify the domain, rather than a series of inflection points which may not be consistently applicable in every case (e.g. pitch 'elbows' are not always identifiable for 'weak' accents [6]).

Looking at the distributions of such features highlights the fact that while indirect responses are often produced with a distinct fall-rise accent, they do not always exhibit the rise part. Perceptual work is currently underway investigating what this rise adds to the interpretation of such dialogue moves. Similarly, we are investigating the role of speaker engagement in the distribution of shape features.

7. References

[1] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proceedings of ICPhS 2007*, 2007, pp. 1065–1068.

[2] V. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging," *Computer Speech & Language*, vol. 23, no. 4, pp. 407–422, 2009.

[3] K. Truong and D. Heylen, "Disambiguating the functions of conversational sounds with prosody: the case of'yeah'," in *Proceedings of Interspeech 2010*. International Speech Communication Association (ISCA), 2010.

[4] M. Steedman, "Information-structural semantics for English intonation," in *Topic and Focus*, C. Lee, M. Gordon, and D. Büring, Eds. Springer, 2007, pp. 245–264.

[5] D. Büring, "On D-Trees, Beans, and B-Accents," *Linguistics & Philosophy*, vol. 26, no. 5, pp. 511–545, 2003.

[6] S. Calhoun, "Information structure and the prosodic structure of English: A probabilistic relationship," Ph.D. dissertation, University of Edinburgh, 2007.

[7] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[8] Y. Xu, "ProsodyPro.praat," http://www.phon.ucl.ac.uk/home/yi/ProsodyPro/, 2005-2011.

[9] M. Liberman and I. Sag, "Prosodic form and discourse function," in *Tenth Regional Meeting, Chicago Linguistic Society*, 1974, pp. 416–427.

[10] M. Gubian, F. Cangemi, and L. Boves, "Automatic and data driven pitch contour manipulation with functional data analysis," in *Proceedings of Speech Prosody 2010, Chicago, USA*, 2010.

[11] C. Lai, "What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue," in *Proceedings of INTERSPEECH'10, Makuhari, Japan, September 2010*, 2010.