UTEP Technical Report UTEP-CS-13-41

Lexical and Prosodic Indicators of Importance in Spoken Dialog

Nigel G. Ward and Karen A. Richart-Ruiz

Department of Computer Science
University of Texas at El Paso
500 West University Avenue
El Paso, TX 79968-0518

email: nigelward@acm.org, karichart@miners.utep.edu

July 23, 2013

This technical report complements the main paper, Patterns of Importance Variation in Spoken Dialog [Ward and Richart-Ruiz, 2013], by providing additional evidence for the claims, additional findings, and more analysis. In particular, we report more on inter-annotator disagreement, on words that correlate with importance, on prosodic features and patterns that correlate with importance, and on how our predictive model of importance might be improved.

Index Terms: annotation, correlations, prediction, prosodic constructions, dialog activity patterns

1 Annotation Procedure and Sources of Disagreements

The main paper overviews the annotation procedure; this section provides some extra details.

After the primary annotator had done 10 minutes of dialog we checked over her work. To do this the second author independently labeled 10 minuts of the same dialogs. then examined the places where her labels differed from the annotator's by more than one point. These disagreements were of four main types:

1. differences due to variation in the placement of boundaries between regions, which arose because the audio was not pre-segmented. One type of difference was variation in the marking of words' exact endpoints. Another type involved within-word variation, for example during words stretched out while the speaker decided how to continue. In such cases the importance seemed to steadily decrease, but the use of discrete labels required the annotators to arbitrarily chose a timepoint where to mark a drop.

	level 0	level 1	level 2	level 3	level 4	level 5	totals
level 0	1072	4	6	18	4	4	1108
level 1	0	2	3	1	1	0	9
level 2	2	1	19	8	3	0	34
level 3	2	0	1	149	46	7	207
level 4	1	0	0	31	313	41	387
level 5	2	0	1	14	69	287	371
totals	1080	7	31	221	437	338	2114

Table 1: Cumulative duration in seconds of regions at each level: annotator 1 by rows, annotator 2 by columns

2. differences arising because the annotator sometimes missed small quiet sounds, especially quiet backchannels that overlapped talk by the main speaker and pre-turn noisy inbreaths.

3. differences in the treatment of repetitions. The annotator tended to ascribe the same importance to both renditions of repeated words, for example in cases of false starts. Logically, one or the other is redundant and thus less important. The second author tended to consider the second rendition more important, as it was generally more fluent and clearer, but one could also argue for the importance of the first rendition, from turn-taking considerations, for example.

4. differences in the ratings of backchannels. Although low in content, these are known to be important for the flow of the dialog.

Given our current level of knowledge, and in particular the lack of any reason to consider our opinions more valid than hers, we chose not to change the labels or procedure. Instead we just sat down with the annotator, discussed the differences, and asked her to pay a little more attention to these aspects for the remainder of the labeling.

The quantitative correspondence between the two sets of labels is shown in Table 1. This is fairly good, since many of the discrepancies are of minor or no real significance, as discussed above, we based the remainder of the analysis on just the labels of the primary annotator.

2 Importance-Correlating Words

As mentioned in the main paper, we examined how words related to importance levels.

Specifically, for each word, we computed the average duration-weighted importance across all occurrences of that word in the annotated data. Table 2 shows the mean and standard deviation of importance for selected words, namely, out of the 100 most frequent words, those with the highest averages (top), with the lowest averages (bottom), and the most frequent (middle). Here the most "frequent" words are not those most frequent over the entire corpus, but over those which overlap regions with a non-zero importance label. Here "words" include also all non-lexical items in the Switchboard transcriptions, including different forms of noise and laughter [ISIP, 2003].

word	mean	std.dev.
everything	4.48	0.69
Texas	4.45	0.76
exercise	4.32	0.82
house	4.29	0.78
...		
that	3.79	0.75
I	3.73	0.78
and	3.72	0.81
uh	3.34	0.89
...		
um	2.99	0.84
um-hum	2.48	0.55
uh-huh	2.46	0.60
(noise)	2.31	1.04

Table 2: Average and Standard Deviation of Importance during Selected Words

One obvious additional factor to consider is context [Ward, 2011]. Table 3 shows words which are predictive and anti-predictive of importance one second later, and Table 4 one second before. Consulting these tables, and the fuller lists, reveals some interesting patterns.

Clearly some words, although not themselves typically very important, are predictors of upcoming importance: for example *because*, *a*, *we*, and *have* all tend to occur with an important region following one second later. (Interestingly *uh* was not an indicator at this offset, despite the fact that it characteristically precedes low-frequency words [Zhu and Penn, 2006].) There are also retrodicting words: *and* and *but* are both typically below average in importance, but both are good indicators that something important was said one second ago. Even more information could surely be obtained by examining word sequences.

Tables 5, 6, and 7 give the corresponding information for words in the interlocutor track. Generally interlocutors don't talk much during and in the vicinity of the other's speaking, so the counts are lower. To avoid reporting means for words that only occured once or twice, here only words among the most frequent 25 are shown.

Several interesting observations can be made. Although *oh* and *uh-huh* are on average low in importance, they are good indicators of something important being said in the other track; and even stronger as retrodictors, along with words like *exactly* and *yes*. On the other hand, *and* is and *but* are good indicators that whatever the other person said a second ago was less important. There are also predictors: *um* and 'vocalized noise' tokens make it more likely that the interlocutor's contribution after a second will be relatively important, and words like *the*, *think* and *really* predict the opposite. These can all be plausibly related to common patterns of turn-taking and interaction. The contributions of interlocutor behavior here illustrate one of the reasons why dialog is often easier to understand than monolog [Branigan et al., 2011], and relate to the relevance for summarization of features relating to listener feedback and discourse cues [Zechner, 2002, Murray et al., 2006].

word	mean	std.dev.
because	4.33	0.74
even	4.25	0.97
exercise	4.18	0.73
a	4.17	0.76
...		
I	3.96	0.80
uh	3.87	0.83
and	3.86	0.80
that	3.85	0.86
...		
car	3.48	0.64
(laughter)	3.33	0.91
guess	3.22	1.08
(noise)	3.17	0.96

Table 3: Average and Standard Deviation of Importance 1 second *after* Selected Words by the Speaker; these words are leading indicators of importance or unimportance in the other track.

word	mean	std.dev.
everything	4.35	1.09
house	4.22	0.86
Texas	4.14	0.14
at	4.08	0.93
...		
and	4.06	0.78
that	3.85	0.80
uh	3.71	0.84
(laughter)	3.69	0.99
...		
right	3.49	1.21
guess	3.46	0.91
yeah	3.29	0.89
(noise)	3.11	1.04

Table 4: Average and Standard Deviation of Importance 1 second *before* Selected Words by the Speaker; these words are lagging indicators of importance or unimportance.

word	mean	std.dev.
oh	4.27	1.00
uh-huh	3.96	0.93
yeah	3.93	0.86
um-hum	3.86	0.90
...		
(laughter)	3.33	0.79
...		
was	3.02	0.87
uh	3.02	0.75
but	3.01	0.53
and	2.97	0.72

Table 5: Average and Standard Deviation of Importance related to Selected Simultaneous Words by the *Interlocutor*. Here *uh-huh*, *yeah* and *um-hum* are among both the most frequent and the most importance-indicating words.

word	mean	std.dev.
(vocalized-noise)	3.80	0.87
um	3.72	1.01
oh	3.70	0.89
yeah	3.68	0.85
um-hum	3.67	0.92
...		
uh-huh	3.51	0.98
(laughter)	3.42	0.75
...		
was	3.22.	0.90
know	3.21	0.85
and	3.09	0.93
the	2.92	0.74

Table 6: Average and Standard Deviation of Importance 1 second *after* Selected Words by the *Interlocutor*; these words are leading indicators of importance or unimportance. *Yeah* and *uh-huh* were among both the most frequent words and the most importance-indicating.

word	mean	std.dev.
exactly	4.46	1.05
oh	4.40	0.78
yes	4.32	0.79
um-hum	4.30	0.67
...		
uh-huh	4.25	0.68
yeah	4.18	0.76
(laughter)	3.63	0.90
...		
so	3.19	0.87
it's	3.18	1.03
(noise)	3.16	1.23
and	3.09	0.90

Table 7: Average and Standard Deviation of Importance 1 second *before* Selected Words by the *Interlocutor*; these words are lagging indicators of importance or unimportance in the other track. *Um-hum* was both one of the most frequent words and one of the most importance-indicating.

3 Importance-Correlating Prosodic Features

The main paper overviews how prosodic features correlate with importance.

These studies were done over 78 prosodic features. These were chosen based on previous investigations of which features are more important for language modeling [Ward et al., 2011, Vega, 2012]. Notably, they are finer-grained near the point of interest. These were also the 78 features used in the predictive models mentioned below.

Tables 8 and 9 give all the correlations and Table 10 shows some highlights, namely the most positively correlating and the most negatively correlating features in each class, if any.

4 Importance-Correlating Dialog Dimensions

The main paper discusses dimensional analysis and overviews what it tells us. One detail not mentioned there is that the dimensions were generated using not the 78 features above, but an older set of 76 features [Ward and Vega, 2012].

Table 11 shows the nine dimensions most strongly related to importance.

Interpretations are taken from [Ward and Vega, 2012], except for dimensions 19 and 75. For these we applied the same methods, interpreting by considering our impressions the dialog state, situation or activity happening at times with extremely high and low values on that dimension, and also considering the raw features that were the strongest contributors to each dimension. Although subjective, we did this analysis with some discipline [Ward and Vega, 2012].

All dimensions turn out to have interpretations which understandably relate to importance, except dimension 75, which we couldn't interpret non-disjunctively (it mostly seems to encode

correlation	feature	offset
0.005	sel-vo	–50 ∼ 0
0.644	sel-vo	–100 ∼ –50
0.642	sel-vo	–200 ∼ –100
0.657	sel-vo	–300 ∼ –200
0.618	sel-vo	–400 ∼ –300
0.578	sel-vo	–800 ∼ –400
0.601	sel-vo	–1600 ∼ –800
0.559	sel-vo	–3200 ∼ –1600
0.519	int-vo	–200 ∼ 0
–0.462	int-vo	–400 ∼ –200
–0.468	int-vo	–800 ∼ –400
–0.498	int-vo	–1600 ∼ –800
–0.505	int-vo	–3200 ∼ –1600
–0.490	sel-ph	–50 ∼ 0
–0.049	sel-ph	–100 ∼ –50
–0.048	sel-ph	–200 ∼ –100
–0.063	sel-ph	–400 ∼ –200
–0.075	sel-ph	–800 ∼ –400
–0.084	int-ph	–200 ∼ 0
0.075	int-ph	–400 ∼ –200
0.068	int-ph	–800 ∼ –400
0.084	sel-pr	–50 ∼ 0
–0.015	sel-pr	–100 ∼ –50
–0.015	sel-pr	–200 ∼ –100
–0.042	sel-pr	–400 ∼ –200
–0.045	sel-pr	–800 ∼ –400
–0.007	int-pr	–200 ∼ 0
–0.032	int-pr	–400 ∼ –200
–0.030	int-pr	–800 ∼ –400
–0.005	sel-sr	–50 ∼ 0
0.273	sel-sr	–100 ∼ –50
0.270	sel-sr	–200 ∼ –100
0.324	sel-sr	–400 ∼ –200
0.364	sel-sr	–800 ∼ –400
0.357	sel-sr	–1600 ∼ –800
0.293	int-sr	–200 ∼ 0
–0.219	int-sr	–400 ∼ –200
–0.216	int-sr	–800 ∼ –400
–0.247	int-sr	–1600 ∼ –800

Table 8: Correlations with importance of all prior-to-frame (past) prosodic features. sel = self, int = interlocutor. vo = volume, ph = pitch height, pr = pitch range, sr = speaking rate proxy. Window start and end times are in milliseconds relative to the frame whose importance is being considered.

correlation	feature	offset
–0.256	sel-vo	+0 ∼ +50
0.637	sel-vo	+50 ∼ +100
0.619	sel-vo	+100 ∼ +200
0.615	sel-vo	+100 ∼ +300
0.641	sel-vo	+300 ∼ +400
0.522	sel-vo	+400 ∼ +800
0.534	sel-vo	+800 ∼ +1600
0.493	sel-vo	+1600 ∼ +3200
0.463	int-vo	+0 ∼ +200
–0.460	int-vo	+200 ∼ +400
–0.459	int-vo	+400 ∼ +800
–0.480	int-vo	+800 ∼ +1600
–0.478	int-vo	+1600 ∼ +3200
–0.451	sel-ph	+0 ∼ +50
–0.050	sel-ph	+50 ∼ +100
–0.051	sel-ph	+100 ∼ +200
–0.073	sel-ph	+200 ∼ +400
–0.100	sel-ph	+400 ∼ +800
–0.113	int-ph	+0 ∼ +200
0.077	int-ph	+200 ∼ +400
0.080	int-ph	+400 ∼ +800
0.100	sel-pr	+0 ∼ +50
–0.015	sel-pr	+50 ∼ +100
–0.015	sel-pr	+100 ∼ +200
–0.042	sel-pr	+200 ∼ +400
–0.043	sel-pr	+400 ∼ +800
–0.007	int-pr	+0 ∼ +200
–0.034	int-pr	+200 ∼ +400
–0.033	int-pr	+400 ∼ +800
–0.008	sel-sr	+0 ∼ +50
0.259	sel-sr	+50 ∼ +100
0.243	sel-sr	+100 ∼ +200
0.284	sel-sr	+200 ∼ +400
0.303	sel-sr	+400 ∼ +800
0.290	sel-sr	+800 ∼ +1600
0.259	int-sr	+0 ∼ +200
–0.226	int-sr	+200 ∼ +400
–0.230	int-sr	+400 ∼ +800
–0.262	int-sr	+800 ∼ +1600

Table 9: Correlations with importance of all after-frame (future) prosodic features, as above.

corr.	feature class	strongest window
0.66	speaker volume	–300 to –200
0.52	interlocutor volume	–200 to 0
0.36	speaker speaking rate	–800 to –400
0.29	interlocutor speaking rate	–200 to 0
0.10	speaker pitch range	0 to 50
0.07	interlocutor pitch height	–400 to –200
–0.01	speaker speaking rate	0 to +50
–0.04	speaker pitch range	–800 to –400
–0.11	interlocutor pitch height	0 to +200
–0.26	speaker volume	0 to +50
–0.26	interlocutor speaking rate	+800 to +1600
–0.34	interlocutor pitch range	+200 to 400
–0.48	interlocutor volume	+1600 to +3200
–0.49	speaker pitch height	–50 to 0

Table 10: Correlations with importance of selected prosodic features, showing for each class of feature the windows over which the correlations were strongest: the most positive and most negative. Times are in milliseconds before (–) or after (+) the start of the 10-millisecond frame whose importance is being predicted.

the brute fact of the existence of a transition from an interlocutor's extended region of high pitch to one by the speaker about 2 seconds in the past).

Dimension 6 matches well with a prosodic construction described in the literature [Ogden, 2012]. Ogden describes a recurring pattern in which a listener expresses agreement with an assessment by producing an upgraded version, for example when one speaker tentatively observes *it's pretty* and the other follows with *absolutely gorgeous* with increased volume, pitch height and pitch range, and "tighter" articulation. This matches exactly what occurs when the value on dimension 6 is high: it is positive to the extent that: the interlocutor was speaking loudly but with low pitch over some time, most strongly around 400 milliseconds in the past, while the speaker was quiet; followed by a loud region by the speaker with a slightly expanded pitch range and increased speaking rate (the upgraded assessment); followed after a short pause by a long and loud continuation by the interlocutor. An example very high on this dimension occurred 309 seconds into dialog sw2402, where A has spoken favorably about warm places:

> A: a lot of people go to Arizona or Florida for the winter and they're able to play all year round but
>
> B: yeah, oh, Arizona's beautiful!

In effect, this analysis led to the rediscovery of a meaningful prosodic pattern, not known to us at the time we did the interpretations. More generally, such analyses may help anchor predictive models in deeper models of the dialog activities or contexts [Lukowicz et al., 2012] at and around the points at which importance happens (or, from the interlocutors' perspective, is accomplished).

correlation	dimension	interpretation
0.44	1	this speaker talking (*vs.* other speaker talking)
0.31	2	both speaking (*vs.* neither speaking)
0.25	7	floor conflict (*vs.* floor sharing)
0.20	8	ending confidently and crisply (*vs.* dragging out a turn)
0.19	3	topic closing (*vs.* topic continuation)
0.18	6	expressing empathy (*vs.* seeking empathy)
0.18	5	turn yield (*vs.* turn grab)
0.17	75	(not interpreted)
0.16	19	solicitous (*vs.* controlling)

Table 11: Interpretations of the dimension directions (positive or negative) correlating most strongly with importance.

	past				future	all
	−1000	−500	−200	0		
speaker	.35	.44	.56	.66	.59	.70
interloc.	.32	.38	.37	.43	.37	.47
both	.41	.45	.62	.71	.65	.74

Table 12: Model Quality, in terms of R^2, as a function of the features used.

5 Prediction Quality and Error Analysis

The main paper gives the key performance results.

Table 12 shows the performance using features only up to various temporal offsets. The rightmost three columns are the same as in Table 2 of the main paper. The leftmost columns are from predictions using cleverly designed featuresets, in which there were fine-grained features up close to the point of prediction. Surprisingly, performance with these clever featuressets was worse than performance using the much smaller featuresets obtained by simply ablating all features which continued on past the point of prediction, as seen in the main paper.

For the 5-training 2-test experiments, is that we chose for the test data dialogs which shared no speakers with the training set. Specifically the training data were tracks sw02055:right, sw02389:right, sw04572:left, sw02436:left and sw02511:right. Sampling every 10 milliseconds, this gave a total of 224495 datapoints. The test data was sw02774:right and sw02442:right, comprising 66262 datapoints.

To judge how significant in practice the prediction errors are, and to look for their causes, we analyzed them in various ways. Figure 1 shows the distributions of the predictor's inferred importance values for regions with each human label. The major problem was a tendency to avoid predicting extreme values, and in particular missing the level-5 labels, as seen also in Figure 1 of the main paper.

To explore more deeply, we went to the audio and listened to regions with a discrepancy between the predicted value and the actual value of 1 or more; specifically all such regions

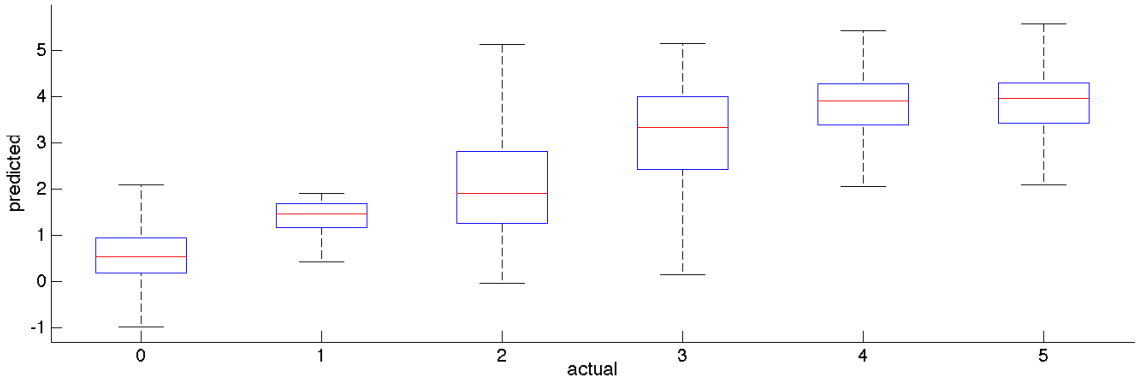

Figure 1: Performance of the linear-regression predictor over all the data. The box-center line is the median, the box edges are the 25th and 75th percentiles, and the region within the whiskers covers 99% of the data.

occurring in 5 minutes of dialog sampled over two conversations. The common causes of errors were, in rough order of frequency and magnitude:

1. poor audio quality, notably inter-track bleeding,
2. differences in boundary assignment, with the predictor often more precise than the annotator in the vicinity of speech-silence boundaries, but sometimes less so, for example anticipating by a few tens of milliseconds the importance transition at an utterance onset,
3. differences during stretched out words, with the predictor sometimes more sensitive to importance variation within a word,
4. the predictor generally gave less importance to the first and last few hundred milliseconds of utterances, including in particular most of most backchannels,
5. the predictor sometimes gave non-zero importance to "hallucinated speech," where a speaker might well have kept talking, or taken the turn, but in fact chose to be silent at that point,
6. the predictor failed to assign enough importance to regions with positional importance, including the summary of a long discussion on a topic, dialog-initial greetings, and dialog closings,
7. the predictor was over-sensitive to volume variation, for example when the speaker apparently was varying the distance from her mouth to the receiver,
8. the predictor responded to the sounds of telephone hang-up at call end, giving them significant importance, but which our annotator left at the default, zero, level,
9. the predictor assigned non-zero importance to some glottalizations (barely-audible occurrences of a few glottal closures that occasionally precede turn- or topic- initiation [Batliner et al., 1993]),

10. the predictor once failed to ascribe adequate importance to a word which locally had no prosodically distinguishing context, but which was the culmination of a rhetorically sophisticated structure and followed a big broad prosodic curve, and finally,

11. there were two "judgment calls": in one the predictor's assigned level-5 importance to the word *colors* in *yes we have brick on the outside and uh it the colors that were there changed*, which upon listening seemed appropriate to us, although the annotator had labeled it a 3 along with all the neighboring words; and in the other the predictor assigned normal importance to an in-passing mention of the speaker's *son*, which however the annotator considered significant, perhaps because she felt any revelation about the family to be semantically interesting and also relevant for rapport-building.

Sometimes speaker differences also appeared to be contributing to some of these discrepancies, as seen also by the fact that some of the correlations between prosodic features and importance differed greatly from speaker to speaker.

This analysis indicates several things.

First, evaluation by measuring correspondences to human annotations can understate the true quality of the model, since many discrepancies, including types 2, 3, 9, and 11, resembled discrepancies between the judgments of the two annotators.

Second, prosody alone is indeed not able to reveal all important regions in dialog. In particular there is a role for longer-context features representing dialog structure and rhetorical structures, to deal with problems 6 and 10, and a role for lexical features and semantic modeling, to deal with problem 11, not to mention semantic importance in task-oriented dialog genres. However, overall, prosody alone gives generally good predictions, with the apparent need for such more complex modeling techniques very limited.

Third, the best way to most improve the current model would be to use better raw prosodic features, including more robust features, better loudness features, and more fine-grained features, to deal with problems 1, 4, 7 and 8. Although likely to have far less impact, use of a model better than linear regression could also improve performance, at least for discrepancies of type 5.

6 Future Work

In addition to the topics for further investigation mentioned in the main paper, several others also appear promising.

A model of importance might be useful not only for spoken dialog. Written language of course lacks prosody, interaction, and the temporal immediacy of spoken dialog, however spoken and textual interactions have many usefully-similar properties [Murray and Carenini, 2008], and models of variance in importance for text also may be worth exploring.

The notion of importance might be clarified by factoring it out into components, including perhaps contentfuless, utility for helping the listener predict what will come next, significance for developing rapport and other interpersonal functions, and indicating to the listener what sort of response would be welcome. We suspect that some of the disagreements between the annotators

may reflect different interpretations of the term importance, which might be representable with different weights for such components.

Acknowledgments

This work was supported in part by the NSF under projects IIS-0914868 and CNS-0837556. We thank Timo Baumann, Alejandro Vega, Shreyas Karkhedkar, Gabriela Almeida and David Novick.

References

[Batliner et al., 1993] Batliner, A., Burger, S., Johne, B., and Kießling, A. (1993). Müsli: a classification scheme for laryngealizations. In *ESCA Workshop on Prosody*, pages 176–179.

[Branigan et al., 2011] Branigan, H. P., Catchpole, C., and Pickering, M. (2011). What makes dialogues easy to understand? *Language and Cognitive Processes*, 26:1667–1686.

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

[Lukowicz et al., 2012] Lukowicz, P., Pentland, A. S., and Ferscha, A. (2012). From context awareness to socially aware computing. *Pervasive Computing, IEEE*, 11(1):32–41.

[Murray and Carenini, 2008] Murray, G. and Carenini, G. (2008). Summarizing spoken and written conversations. In *Proc. Empirical Methods in Natural Language Processing*, pages 773–782.

[Murray et al., 2006] Murray, G., Renals, S., Carletta, J., and Moore, J. (2006). Incorporating speaker and discourse features into speech summarization. In *Proc. HLT-NAACL*, pages 367–374. Association for Computational Linguistics.

[Ogden, 2012] Ogden, R. (2012). Prosodies in conversation. In Niebuhr, O., editor, *Understanding Prosody: The role of context, function, and communication*, pages 201–217. De Gruyter.

[Vega, 2012] Vega, A. (2012). On the selection of prosodic features for language modeling. University of Texas at El Paso, Computer Science Department Masters Thesis.

[Ward, 2011] Ward, N. G. (2011). Temporal distributional analysis. In *SemDial*.

[Ward and Richart-Ruiz, 2013] Ward, N. G. and Richart-Ruiz, K. A. (2013). Patterns of importance variation in spoken dialog. In *14th SigDial*.

[Ward and Vega, 2012] Ward, N. G. and Vega, A. (2012). A bottom-up exploration of the dimensions of dialog state in spoken interaction. In *13th Annual SIGdial Meeting on Discourse and Dialogue*.

[Ward et al., 2011] Ward, N. G., Vega, A., and Baumann, T. (2011). Prosodic and temporal features for language modeling for dialog. *Speech Communication*, 54:161–174.

[Zechner, 2002] Zechner, K. (2002). Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28:447–485.

[Zhu and Penn, 2006] Zhu, X. and Penn, G. (2006). Summarization of spontaneous conversations. In *Proc. of Interspeech*, pages 1531–1534.