Patterns of Importance Variation in Spoken Dialog

Nigel G. Ward
University of Texas at El Paso
El Paso, Texas, 79968 USA
nigelward@acm.org

Karen A. Richart-Ruiz
University of Texas at El Paso
El Paso, Texas, 79968 USA
karichart@miners.utep.edu

Abstract

Some things people say are more important, and some less so. Importance varies from moment to moment in spoken dialog, and contextual prosodic features and patterns signal this. A simple linear regression model over such features gave estimates that correlated well, 0.83, with human importance judgments.

1 Importance in Language and Dialog

Not everything people say to each other is equally important, for example many *um*s and *uh*s have almost no significance, in comparison to those content words or nuances that are critical in one way or another.

Many language processing applications need to detect what is important in the input stream, including dialog systems and systems for summarization, information retrieval, information extraction, and so on. Today this is primarily done using task-specific heuristics, such as discarding stopwords, giving more weight to low frequency words, or favoring utterances with high average pitch. In this paper, however, we explore a general, task-independent notion of importance, taking a dialog perspective.

Section 2 explains our empirical approach. Sections 3 and 4 explore the individual prosodic features and longer prosodic patterns that dialog participants use to signal to each other what is important and unimportant. Section 5 describes predictive models that use this information to automatically estimate importance and Section 6 summarizes the significance and future work needed.

2 Annotating Importance

No standard definition of importance is useful for describing what happens, moment-by-moment, in spoken dialog. The closest contender would be entropy, as defined in information theory. For text we can measure the difficulty of guessing letters or words, as a measure of their unpredictability and thus informativeness (Shannon, 1951), but this is indirect, time-consuming, and impossible to apply to non-symbolic aspects of language. We can also measure the value of certain information, such as prosody, for improving the accuracy of predictions, but again this is indirect and time-consuming (Ward and Walker, 2009).

We therefore chose to do an empirical study. We hired a student to annotate importance. Wanting to capture her naive judgments, atheoretically, we did not precisely define importance for her. Instead we discussed the concept briefly, noting that importance may be judged: not just by content but also by value for directing the future course of the dialog, not just from the speaker's perspective but also from the listener's, and not just from the words said but also from how they were said.

The labeling tool used enabled the annotator to navigate back and forth in the dialogs, listen to the speakers together in stereo or independently, delimit regions of any desired size including words and word fragments, and ascribe to each region an importance value. While importance is continuous, for convenience we used the whole numbers from 0 to 5, with 5 indicating highest importance, 4 typical importance, 3 somewhat less importance, 2 and 1 even less, and 0 silence. To have a variety of speakers, topics, and speaking styles, the material was from the Switchboard corpus (Godfrey et al., 1992).

Proceedings of the SIGDIAL 2013 Conference, pages 107–111,
Metz, France, 22-24 August 2013.

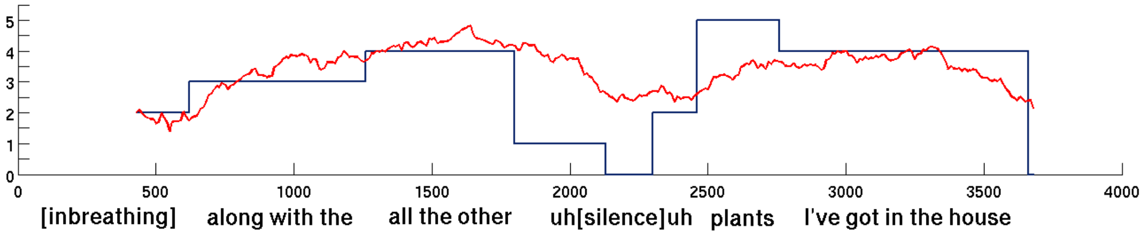

Figure 1: Importance versus Time, in milliseconds. Rectangular line: Annotator judgments; Jagged line: Predictions (discussed below). The words are all by one speaker, horizontally positioned by approximate occurrence.

In total, she labeled both tracks of just over 100 minutes of dialog. There was diversity in labels, supporting our belief that importance is not monotone: the largest fraction of non-zero-labeled regions, covering 38% of the total time, was at level 4, but there were also 20% at level 3 and 37% at level 5. In general importance was variable, on average staying at the same level for only 1.5 seconds. Figure 1 illustrates.

In parallel, the second author labeled 17 minutes of the same dialogs[1]. The agreement in terms of Kappa was .80 ("very good") across all categories, and .67 ("good") excluding the zero-level labels, which were mostly for silent regions and thus easy to agree on. In terms of Weighted Kappa, appropriate here since the labels are ordered (and thus, for example, a 1-point difference matters much less than a 5-point difference), the agreement levels were .92 and .71, for all and for the zero-excluding sets, respectively. The differences were mainly due to minor variations in boundary placement, missing labels for small quiet sounds such as inbreaths and quiet overlapping backchannels, and different ratings of repeated words, and of backchannels (Ward and Richart-Ruiz, 2013).

3 Correlating Prosodic Factors

First we briefly examined lexical correlates of importance, by examining the average importance of words in this corpus (Ward and Richart-Ruiz, 2013). To summarize some key findings: Less frequent words tend to have higher average per-word importance, however ratings vary widely, depending on context. Some words have effects at a distance, for example, *because* tends to indicate that whatever is said one second later will be important. The interlocutor's words can also be informative, for example *oh* and *uh-huh* tend to indicate that whatever the interlocutor said one second ago was important. The "words" with the most extreme average importance — notably *uh-huh*, *um-hum*, *um* and laughter — are fillers, backchannels and other vocalizations of types which can be detected well from the prosodic and interactional contexts (Neiberg and Gustafson, 2011; Truong and van Leeuwen, 2007). Thus a word-based model of importance would be challenging to build and might not have much value. We therefore turned our attention to prosody.

While prosody-importance connections have not been considered directly, several studies have found correlations between prosodic features and various importance-related constructs, such as predictability, involvement, engagement, activation, newness, and interest (Bell et al., 2009; Yu et al., 2004; Batliner et al., 2011; Roehr and Baumann, 2010; Oertel et al., 2011; Hsiao et al., 2012; Kahn and Arnold, 2012; Kawahara et al., 2010). However these studies have all been limited to specific features, functions, or hypotheses. Our aims being instead exploratory, we looked for features, from among a broad inventory, which correlate with importance, as it occurs in a broad variety of contexts.

Our feature inventory included features of 8 classes: four basic types — volume, pitch height, pitch range, and speaking-rate — each computed for both participants: the speaker and the interlocutor. Within each class, features were computed over windows of various widths and at various offsets, for a total of 78 features (Ward and Richart-Ruiz, 2013).

[1] All labels are freely available at http://www.cs.utep.edu/nigel/importance/

The speaker features correlating most strongly with importance were volume and speaking rate. Although the very strongest correlations were with volume slightly in the past, volume both before and after the current moment was strongly correlated over all windows, with one exception. Speaker pitch height, in contrast, correlated negatively with importance across all windows, contrary to what is often seen in monolog data.

The interlocutor features correlating most strongly with importance were again volume and speaking rate, but only over windows close to the point of interest, perhaps due to co-construction or supportive back-channeling; over more distant windows, both past and future, these correlate negatively. Interlocutor pitch range correlated negatively over all windows.

4 Correlating Dialog-Activity Patterns

Thus we find that some prosodic features have different effects depending on their offset from the frame of interest. Perhaps prosody is not just marking importance vaguely somewhere in the area, but more precisely indicating important and unimportant moments.

To explore this we used Principal Components Analysis (PCA), as described in detail in (Ward and Vega, 2012). In short, this method finds patterns of prosodic features which co-occur frequently in the data, and so provides an unsupervised way to discover the latent structure underlying the observed regularities. We correlated the dimensions resulting with PCA with the importance values. Many dimensions had significant correlations, indicating that importance relates to many prosodic structures and contexts. Each dimension had two characteristic patterns, one corresponding to high values on that dimension and one to low values. We were able to interpret most of these in terms of dialog activities (Ward and Vega, 2012).

Tending to be more important was: speech in the middle of other speech (dimension 1), rather than words snuck in while the other has the floor; simultaneous speech (dimension 2), understandably as such times tended to be high in involvement and/or backchannels; times of encountering and resolving turn conflicts (dimension 7), more than places where the participants were supportively interleaving turns, which in this corpus were generally more phatic than contentful; crisp turn ends (dimension 8), rather than slow repetitious wind-downs; "upgraded assessments," in which a speaker agrees emphatically with an assessment made by the other (dimension 6); and times when speakers were solicitous, rather than controlling (dimension 19). Dimension 6 is interesting in that it matches an interaction pattern described as an exemplar of prosodic co-construction (Ogden, 2012). Dimension 19 was one of those underlying the exception noted above: the negative correlation between importance and speaker volume over the window from 0–50 milliseconds after the point of prediction. Upon examination, low volume at this offset often occurred when seeking agreement and during quiet filled pauses in the vicinity of high-content words.

model	correlation	m.a.e.
m5pTree decision tree	.38	1.21
neural network	.66	1.20
simple linear regression	.79	.89
linear regression	.83	.75
ditto, past-only features	.83	.79

Table 1: Prediction Quality in terms of correlation and mean absolute error, for various learning algorithms.

5 Predictive Models

We next set out to build predictive models, for two reasons: to judge whether the features discussed above are adequate for building useful models, and to determine what additional factors would be required in a more complete model.

The task is, given a timepoint in a track in a dialog, to predict the importance of what the speaker is saying at that moment. Our performance metrics were the mean absolute error and the correlation coefficient, computed over all frames; thus a predictor is better to the extent that its predictions are close to and correlate highly with the annotator's labels, including the implicit zero labels in regions of silence or noise.

We built models using four algorithms in Weka. All models performed poorly on dialogs for which there was cross-track bleeding or other noise. As these are artifacts of this corpus and would not be relevant for most applications, our main evaluation used only the five tracks with good audio quality. These all had different speakers. We did five-fold cross-validation on this; Table 1 gives the results. Linear regression was best, by both measures and

	past			future	all
	−400	−200	0		
speaker	.55	.64	.66	.59	.70
interloc.	.37	.43	.43	.37	.47
both	.62	.70	.71	.65	.74

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

across every fold, and this was consistent for all the other training and test sets tried.

To compare the performance of this predictor to human performance, we also trained a model using 5 tracks to predict performance over two test tracks, a total of 224495 test datapoints, which the second judge also had annotated. Over these the predictor did almost as well as second judge in correlation (.88 versus .92), but not so well in terms of mean absolute error (.75 versus .31).

Analyzing the errors, we noted several types of cause (Ward and Richart-Ruiz, 2013). First, performance varied widely across tracks, with mean absolute errors from .55 to .97, even though all the features were speaker-normalized. The high value was for a speaker who was an outlier in two respects: the only female among four males, and the only East-Coast speaker among four Texans. Thus results might be improved by separately modeling different genders and dialects. Second, predictions were often off in situations like those where the two human judges disagreed. Third, most of the errors were due to feature-set issues: robustness, poor loudness features, and not enough fine-grained features. Fourth, our prosodic-feature-only model did very poorly at distinguishing between the highest importance levels, 4 and 5, but was otherwise generally good.

Table 2 shows how performance varies with the features used; here quality is measured using simply the R^2 of a linear regression over all the data. Performance is lower with only the left-context features, as would be required for real-time applications, but not drastically so; as seen also in the last line of Table 1. Performance is only slightly lower when predicting slightly in advance, without using any features closere than 200 ms prior to the prediction point, but notably worse 400 ms before. Features of the interlocutor's behavior are helpful, partially why explaining dialog can be easier to understand than monolog (Branigan et al., 2011).

6 Broader Significance and Future Work

Sperber and Wilson argue that "attention and thought processes . . . automatically turn toward information that seems relevant: that is, capable of yielding cognitive effects" (Sperber and Wilson, 1987). This paper has identified some of the cues that systems can use to "automatically turn toward" the most important parts of the input stream. Overall, these findings show that task-independent importance can be identified fairly reliably, and that it can be predicted fairly well using simple prosodic features and a simple model. Significantly, we find that importance is frequently not signaled or determined by one participant alone, but is often truly a dialog phenomenon. We see three main directions for future work:

First, there is ample scope to build better models of importance, not only by pursuing the prosodic-feature improvements noted above, but in examining lexical, semantic, rhetorical-structure and dialog-structure correlates of importance.

Second, one could work to put our pretheoretical notion of importance on a firmer footing, perhaps by relating it to entropy, or to the time course of the psychological processes involved in retrieving, creating, managing, and packaging information into speech; or to the design and timing of dialog contributions so as not to overload the listener's processing capacity.

Third, there are applications. For example, a dialog system needing to definitely convey some information to the user could use an appropriate prosodic lead-in to signal it properly, doing an interactional dance (Gratch et al., 2007; Brennan et al., 2010) to prepare the recipient to be maximally receptive at the moment when the critical word is said. Another potential application is in voice codecs, as used in telecommunications. Today's codecs treat all speech as equally valuable. Instead we would like to transmit more important words and sounds at higher quality, and less important ones at lower quality, thereby increasing perceived call quality without increasing the average datarate, of course while properly considering all perceptual factors (Voran and Catellier, 2013).

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

Anton Batliner, Stefan Steidl, Bjorn Schuller, et al. 2011. Whodunnit: Searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech and Language*, 25:4–28.

Alan Bell, Jason M. Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60:92–111.

Holly P. Branigan, C.M. Catchpole, and M.J. Pickering. 2011. What makes dialogues easy to understand? *Language and Cognitive Processes*, 26:1667–1686.

Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. 2010. Two minds, one dialog: Coordinating speaking and understanding. In Brian H. Ross, editor, *The Psychology of Learning and Motivation, volume 53*, pages 301–344. Elsevier.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520.

Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, pages 125–138. Springer.

Joey Chiao-yin Hsiao, Wan-rong Jih, and Jane Yung-jen Hsu. 2012. Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns. In *Activity Context Representation Workshop at AAAI*.

Jason M. Kahn and Jennifer E. Arnold. 2012. A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language*, 67:311–325.

Tatsuya Kawahara, K.Sumi, Z.Q. Chang, and K.Takanashi. 2010. Detection of hot spots in poster conversations based on reactive tokens of audience. In *Interspeech*, pages 3042–3045.

Daniel Neiberg and Joakim Gustafson. 2011. A dual channel coupled decoder for fillers and feedback. In *Interspeech 2011*, pages 3097–3100.

Catharine Oertel, Stefan Scherer, and Nick Campbell. 2011. On the use of multimodal cues for the prediction of degrees of involvment in spontaneous conversation. In *Interspeech*.

Richard Ogden. 2012. Prosodies in conversation. In Oliver Niebuhr, editor, *Understanding Prosody: The role of context, function, and communication*, pages 201–217. De Gruyter.

Christine Tanja Roehr and Stefan Baumann. 2010. Prosodic marking of information status in German. In *Speech Prosody Conference*.

Claude E. Shannon. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64.

Dan Sperber and Deirdre Wilson. 1987. Précis of Relevance: Communication and cognition. *Behavioral and Brain Sciences*, 10(04):697–710.

Khiet P. Truong and David A. van Leeuwen. 2007. Automatic discrimination between laughter and speech. *Speech Communication*, 49:144–158.

Stephen D. Voran and Andrew A. Catellier. 2013. When should a speech coding quality increase be allowed within a talk-spurt? In *IEEE ICASSP*.

Nigel G. Ward and Karen A. Richart-Ruiz. 2013. Lexical and prosodic indicators of importance in spoken dialog. Technical Report UTEP-CS-13-41, University of Texas at El Paso, Department of Computer Science.

Nigel G. Ward and Alejandro Vega. 2012. A bottom-up exploration of the dimensions of dialog state in spoken interaction. In *13th Annual SIGdial Meeting on Discourse and Dialogue*.

Nigel G. Ward and Benjamin H. Walker. 2009. Estimating the potential of signal and interlocutor-track information for language modeling. In *Interspeech*, pages 160–163.

Chen Yu, Paul M. Aoki, and Alison Woodruff. 2004. Detecting user engagement in everyday conversations. In *Interspeech*, pages 1329–1332.