Towards Empirical Dialog-State Modeling and its Use in Language Modeling

Nigel G. Ward, Alejandro Vega

Department of Computer Science, University of Texas at El Paso, El Paso, Texas, United States

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

Abstract

Inspired by the goal of modeling the dialog state and the speaker's mental state, moment by moment, we apply Principal Component Analysis to a vector of 76 prosodic features spanning 6 seconds of context. This gives a multidimensional representation of the current state. We find that word probabilities vary strongly with several of these dimensions, that the use of this information in a language model gives a 27% reduction in perplexity, and that many of the dimensions do relate to aspects of mental state and dialog state.

Index Terms: prosody, context, principal component analysis, perplexity, dimensions, dialog activities

1. Mind Models, Dialog States, Language Modeling and Prosody

Speech processing systems should ideally incorporate models of the speaker's mental processes [1]: we would like to be able to pinpoint what is in the speaker's mind at any moment. If we could even approach this, speech recognizers could use this information to great benefit. Today, however, this ideal is distant.

Nevertheless there is a sense in which dialog systems, at least, do this: they model the speaker as being in one of a finite set of dialog states (for example listening, giving a yes/no answer, saying a number, asking a question, and so on), or as having one of a finite set of goals, and possibly represent the current state as a probability distribution over such a set of states [2, 3, 4].

However we potentially could do much better. One approach is to work more domain-independent information into the state representations, as so much of what happens in speech is not directly about domain content. This has been done in particular for non-propositional aspects, such as confidence and affect, as surveyed in [5]; and general models and methods also exist [6, 7, 8]. However these have not been applied to speech recognition. For this we especially want models which are both empirically grounded and high in temporal resolution. For this purpose, models using prosodic information seem promising, as prosody strongly reflects dialog state, cognitive state, attitude, and related functions. For example, it is known that slow speaking rate and flat pitch characterize disfluent regions and processes of lexical access, and that quiet regions are associated with, among other things, expressions about veracity and evidence [9, 10].

While prosodic context information has been applied to language modeling [9, 11], the benefits obtained have fallen short of expectations. We here explore a new method which may do better in that it is:

Empirical Rather than using aspects of state deemed interesting a priori or taken from existing taxonomies, the dialog aspects modeled are those discovered to be important in the data.

Broad Rather than using prosodic features computed over small time intervals (e.g. words), it uses features that cover 6-second regions.

Rich Rather than representing only a few dimensions of dialog state, it includes dozens.

Scalar Rather than representing dialog states as discrete, or composed of discrete-valued factors, it represents all aspects of dialog state with continuous scales.

Continuous Rather than modeling state for utterance-sized units, it models it at every instant in time.

2. Decorrelating Prosodic Features

In addition to the ambitious, high-level motivation presented above, this work also has a practical motivation. In previous uses of prosody in language modeling [11], we found that conditioning the probability estimates of words on individual contextual features was promising, but that combined models with multiple features gave less-than-additive perplexity benefits. Using the method of examining specific words where the combined model did poorly [12], we found that many arose because the features were largely redundant — for example, volume, pitch height, and pitch range all correlate — but our modeling method incorporated an independence assumption. While not a terrible problem for four-feature models, this looked likely to prevent the effective use of wider prosodic contexts with more features.

There are various ways to deal with large messy sets of prosodic features. If the ultimate use can be formulated

as a classification problem, that is a supervised learning problem, then non-linear classifiers and/or feature selection algorithms [13] can be effective, but the language modeling problem is hard to formulate in this way. Another approach is manually seek for a set of underlying prosodic features that independently contribute to the meaning or state [14], but linguistic studies of prosody have shown how difficult this is. Another approach is to model the typical prosodic contexts of words using a Gaussian Mixture Models, however this has not so far been successful [15].

We therefore chose to try Principal Components Analysis (PCA), to go from the raw, messily correlated features to a "latent" orthogonal set of components encoding the same information. While PCA has previously been applied to prosodic features, as discussed in [16] and the references therein, it has not before been used in language modeling.

3. The Features, the Data, and Principal Component Analysis

Our feature set consisted of 76 prosodic features, taken from both the immediate past and the immediate future, as dialog state relates to both; since current state depends on previous information and it can predict future actions. We used features from both speaker and interlocutor, as the dialog state at each moment depends on both the speaker's actions and those of his or her interlocutor.

Thinking that fairly short-term dialog states and mental states would be most informative about word identity, we used features over 0–3 seconds before word onset and 0–3 seconds after, for the word in question. We thus included prosodic information from the word itself, although for recognition purposes such information probably belongs more in the acoustic model than in the language model.

The specific features were derived from a basic set of four that have proven useful for language modeling [11]: speaking rate, volume, pitch height, and pitch range. All features are speaker normalized. Details are given in [16].

Using the Switchboard corpus, a large collection of spontaneous dialogs between two strangers over the telephone [17, 18], we collected datapoints from both sides of 20 dialogs, totaling about 2 hours of conversation, sampling at 100ms intervals, giving a total of 600,000 datapoints. We then applied standard PCA, obtaining 76 new features, referred to below as PC 1 through PC 76, ordered by how much of the variance they explained. The top 20 components explained 81% of the variance.

4. Use in Language Models

Expecting different words to be more common in different prosodic contexts, we characterized each context by the value of that context on each component. For example, if an unknown word in the test set occurred in a prosodic context where the value on PC 1 is high, then we expect it to be more likely to be a word frequently observed in the training set in contexts where the value of PC 1 was high.

We therefore computed the frequency of every word in every context. Contexts for each principal component are in terms of four regions, namely the quartiles computed from the distribution of values for that component. From this we compute for each word and each context a scaling factor representing our estimate of how much more or less likely than usual that word is in that context, as detailed in [11]. To apply this information we combine it with a baseline language model by multiplying, after raising the scaling factor to the kth power to weight it appropriately. In the equation, P_S is the final probability estimate for word w_i in lexical context c and prosodic context x, derived from the baseline probability estimate P_b, the scaling factor S, and the weight k.

$$P_s(w_i|c,x) = P_b(w_i|c) * S(w_i|x)^k \quad (1)$$

We built our models using the Switchboard corpus. First we built a baseline trigram backoff model using default SRILM parameters on 90 hours of dialog [11, 19]; this had a perplexity of 111.36 on the test data. We then built 76 individual language models using the same data, each combining the baseline with the scaling factors from one component. These were all evaluated with weight k equal to 0.3.

Finally we built a combined language model using the best 25 components: these were made by simply included more sets of scaling factors, all multiplied together. Weights k for the final combined model were optimized using a separate subset of the data for tuning. The test set consisted of 45 tracks from Switchboard, totaling about and 225 minutes of spoken dialog. Details of the evaluation method appear elsewhere [11].

5. Results

Table 1 presents the perplexity benefits for the 15 principal components which gave the greatest reductions. Interestingly these were not all among the top components: apparently the ability to explain much of the variance does not much correlate with the ability to provide useful information to a language model.

When used in combination, with default weights ($k = 0.3$), the top 25 principal component models provided a 21.7% perplexity reduction, not far below that predicted from the individual perplexities benefits, 24.8%. Using weights obtained by tuning, a perplexity benefit of 26.8% was obtained, as shown in Table 2. This final model improved the estimates for 66% (15766/23836) of the words in the test set.

Model	Perplexity Reduction	Weight in the Tuned Combined Model	Interpretation
PC 12	4.1%	.70	claiming the floor vs. yielding the floor
PC 62	3.4%	.55	explaining/excusing oneself vs. blaming someone/something
PC 72	2.3%	.45	fast and sloppy vs. thoughtful and clear
PC 25	1.4%	.50	personal experience vs. second-hand opinion
PC 15	1.1%	.50	speaking before ready vs. presenting held-back information
PC 13	1.1%	.60	starting a contrasting statement vs. starting a restatement
PC 21	1.0%	.50	mitigating a potential face threat vs. agreeing with an amusing anecdote
PC 30	1.0%	.50	saying something predictable vs. prepare to take a new tack
PC 1	1.0%	.25	this speaker talking vs. other speaker talking
PC 10	0.9%	.25	engaged in lexical access / memory retrieval vs. disengaged from dialog
PC 23	0.9%	.60	closing out a topic vs. starting or continuing one
PC 6	0.9%	.50	seeking empathy vs. expressing empathy
PC 26	0.9%	.35	signaling interestingness vs. downplaying the current information
PC 24	0.9%	.45	agreeing and preparing to move on vs. jointly focusing
PC 18	0.9%	.45	seeking sympathy vs. expressing sympathy

Table 1: The top 15 components, perplexity reductions relative to the baseline obtained with each individually (with weight $k = 0.3$), weights k used in the combined model, and dialog/pragmatic/cognitive interpretations of the components.

Model	Perplexity Reduction
25 components, default weights	21.7%
25 components, tuned weights	26.8%

Table 2: Perplexity Benefits with Combined Models. The 25 components include those in Table 1 plus ten more

6. The Informative Aspects of Dialog State

Wanting to understand better how this method works, we examined some of the principal components and the words they favored. To try to interpret each component we considered four sources of information: the words occurring at times with extremely high and low values on that component; our impressions the dialog state, situation or activity happening at those times; the raw features that were the strongest factors in that component; and the words most characteristic of each of the quartiles of that component. Although subjective, we did this analysis with some discipline [16]. For each component, it was the listening to extreme cases that gave the most coherent picture, with the other kinds of evidence reinforcing or being at least compatible with that interpretation.

For example, for the most valuable principal component, PC 12, timepoints with low values on this dimension generally seemed to be staking a claim to the floor, revealing the intention to talk on for several seconds, sometimes as topic resumptions. Points with high were generally floor yields, and sometimes sounded negative or distancing. The most characteristic words (those with the highest and lowest scaling factors) for the lowest and highest quartiles are shown in Table 3. Slow future speaking rate, by both speakers, aligned with the low values, and fast rate with the high values. Overall we identify this dimension with a floor claim/yield continuum.

Looking again at Table 3, some of the words that are common in each quartile fit well with our interpretation of the dimension. For example, in the lowest quartile, *ohh* and *realize* do plausibly seem likely when the speaker is claiming the floor. Others, however, might not be expected, a priori, to fit this description, for example *realize*, or *while*. However there is ample corpus-based evidence that words correlate with various emotional and social functions in unsuspected ways [20]. For example, in writing, it has been found that in emotional passages articles and prepositions are less frequent than usual, whereas pronouns, auxiliary verbs, and negations are more frequent than usual. Thus the characteristic and uncharacteristic of words in the various dimensions may reflect patterns that could lead to similar psychometric findings.

Turning now to the second most valuable component, PC 62, low values frequently appear at times where the speaker is justifying him- or herself, for example in not exercising, or in liking a certain often-disparaged football team. At these times the speaker tends to have low pitch and volume and a steady speaking rate. High values occur frequently at times when the speaker is blaming or deploring someone or something, for example a team that fails to win, or a garden full of weeds. These times are tend to have high intensity and pitch. Words common at low points include *upset, ago, extremely*, and *especially*; words common at high points include *background, factor, husband, technology* and *politicians*.

Looking at the two next most valuable components: PC 72 represents the continuum between fast, low con-

	High Scaling-Factor Words	Low Scaling-Factor Words
Lowest Quartile	ohh, reunion, realize, hands, oil, hearing, while, language, anywhere, long ...	... company, difference, fifteen, computer
Highest Quartile	quickly, puppy, tickets, picked, expect, technology, countries, company, fifteen, companies ...	... news, ooh, anywhere, realize

Table 3: Most characteristic and uncharacteristic words for two quartiles of component 12.

tent and poorly articulated speech and careful, thoughtful, clear speech; and PC 25 speaking from second-hand knowledge and/or listing evidence versus speaking from personal knowledge and unique experience. However not all dimensions had clear dialog-related interpretations; for example PC 29, the 18th most beneficial component, appears to be describable best as relating simply to the presence of a stressed word. The right column of Table 1 summarizes the interpretations for the 15 most beneficial components; evidence for some of these interpretatations is given elsewhere [16].

Interestingly some of the principal components which explained most of the variance were not among those which had value for language modeling. In particular, the top components relating to turn-taking (PCs 2, 5, 7, and 8), to topic change (3 and 9), and grounding (4) were not very informative regarding word occurrences. This might be because the trigrams already provided enough information about the likely words in such contexts or perhaps because in English these functions are handled primarily by prosody without help from words.

7. Conclusions

Principal Component Analysis enables the effective use of contextual prosodic information for language modeling. This technique requires no hand labeling and should be easy to apply to new corpora and domains.

Moreover many of the dimensions given by this technique relate to mental states and dialog states. This suggests that the goal of modeling dialog state and speaker mental state is not pie in the sky but a reasonable near-term objective for speech science and engineering.

8. Acknowledgments

This work was supported in part by NSF Award IIS-0914868. We thank Olac Fuentes, Justin McManus and Shreyas Karkhedkar.

9. References

[1] H. Fujisaki, "In search of models in speech communication research," in *Interspeech*, pp. 1–10, 2008.

[2] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. T. R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, pp. 339–373, 2000.

[3] S. Young, "Using POMDPs for dialog management," in *IEEE/ACL Workshop on Spoken Language Technology (SLT 2006)*, 2006.

[4] A. Raux, N. Mehta, D. Ramachandran, and R. Gupta, "Dynamic language modeling using bayesian networks for spoken dialog systems," in *Interspeech*, pp. 3030–3033, 2010.

[5] J. C. Acosta and N. G. Ward, "Achieving rapport with turn-by-turn, user-responsive emotional coloring," *Speech Communication*, vol. 53, pp. 1137–1148, 2011.

[6] O. Lemon, L. Cavedon, and B. Kelly, "Managing dialogue interaction: A multi-layered approach," in *4th (ACL) SIGdial Workshop on Discourse and Dialogue*, 2003.

[7] J. R. Tetreault and D. J. Litman, "Using reinforcement learning to build a better model of dialogue state," in *EACL*, 2006.

[8] H. Bunt, "Multifunctionality in dialogue," *Computer Speech and Language*, vol. 25, pp. 222–245, 2011.

[9] A. Stolcke, E. Shriberg, D. Hakkani-Tur, and G. Tur, "Modeling the prosody of hidden events for improved word recognition," in *Proceedings of the 6th European Conference on Speech Communication and Technology*, 1999.

[10] 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.

[11] N. G. Ward, A. Vega, and T. Baumann, "Prosodic and temporal features for language modeling for dialog," *Speech Communication*, vol. 54, pp. 161–174, 2011.

[12] N. G. Ward, A. Vega, and D. G. Novick, "Lexico-prosodic anomalies in dialog," in *Speech Prosody*, 2010.

[13] A. Batliner, S. Steidl, B. Schuller, *et al.*, "Whodunnit: Searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech and Language*, vol. 25, pp. 4–28, 2011.

[14] D. R. Ladd, K. E. A. Silverman, F. Tokmitt, *et al.*, "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect," *Journal of the Acoustic Society of America*, vol. 78, pp. 435–444, 1985.

[15] S. A. Karkhedkar and N. G. Ward, "Representing the prosodic context of words using Gaussian mixture models," in *Speech Prosody 2012*, 2012.

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

[17] 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.

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

[19] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. Intl. Conf. on Spoken Language Processing*, 2002.

[20] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, pp. 24–54, 2010.