A Bottom-Up Exploration of the Dimensions of Dialog State in Spoken Interaction

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

Alejandro Vega
Department of Computer Science
University of Texas at El Paso
El Paso, Texas, 79968, USA
avega5@miners.utep.edu

Abstract

Models of dialog state are important, both scientifically and practically, but today's best build strongly on tradition. This paper presents a new way to identify the important dimensions of dialog state, more bottom-up and empirical than previous approaches. Specifically, we applied Principal Component Analysis to a large number of low-level prosodic features to find the most important dimensions of variation. The top 20 out of 76 dimensions accounted for 81% of the variance, and each of these dimensions clearly related to dialog states and activities, including turn taking, topic structure, grounding, empathy, cognitive processes, attitude and rhetorical structure.

1 Introduction

What set of things should a dialog manager be responsible for? In other words, which aspects of the current dialog state should the dialog manager track?

These questions are fundamental: they define the field of computational dialog modeling and determine the basic architectures of our dialog systems. However the answers common in the field today arise largely from tradition, rooted in the concerns of precursor fields such as linguistics and artificial intelligence (Traum and Larsson, 2003; McGlashan et al., 2010; Bunt, 2011).

We wish to provide a new perspective on these fundamental questions, baed on a bottom-up, empirical investigations of dialog state. We hope thereby to discover new facets of dialog state and to obtain estimates of which aspects of dialog state are most important.

2 Aims

There are many ways to describe dialog state, but in this paper we seek a model with 7 properties:

Orthogonal to Content. While the automatic discovery of content-related dialog states has seen significant advances, we are interested here in the more general aspects of dialog state, those that occur across many if not all domains.

Scalar. While it is descriptively convenient to refer to discrete states (is-talking, is-waiting-for-a-yes-no-answer, and so on), especially for human analysts, in general it seems that scales are more natural for many or all aspects of dialog state, for example, one's degree of confidence, the strength of desire to take the turn, or the solidity of grounding.

Non-Redundant. While various levels and angles are used in describing aspects of dialog state — and many of these are interrelated, correlated, and generally tangled — we would like a set of dimensions which is as concise as possible and mutually orthogonal.

Continuously Varying. While it is common to label dialog states only at locally stable times, for example when neither party is speaking, or only over long spans, for example, utterances, we want a model that can support incremental dialog systems, able to describe the instantaneous state at any point in time, even in the middle of an utterance.

Short-Term. While aspects of dialog state can involve quite distant context, we here focus on the aspects important in keeping the dialog flowing over

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 198–206,
Seoul, South Korea, 5-6 July 2012. ©2012 Association for Computational Linguistics

short time-scales.

Non-Exhaustive. While dialog states can be arbitrarily complex, highly specific, and intricately related to content, a general model can only be expected to describe the frequently important aspects of state.

Prioritized. While no aspects of dialog are uninteresting, we want to know which aspects of dialog state are more important and commonly relevant.

3 Approach

To be as empirical as possible, we want to consider as much data as possible. We accordingly needed to use automatic techniques. In particular, we chose to base our analysis on objective manifestations of dialog state. Among the many possible such manifestations — discourses markers, gesture, gaze, and so on — we chose to use only prosody. This is because the importance of prosody in meta-communication and dialog control has often been noted, because the continuous nature of (most) prosodic features is convenient for our aims, and because prosodic features are relatively easy to compute.

Given our aims and such features, it is natural to do Principal Components Analysis (PCA). This well-known method automatically identifies the factors underlying the observed variations across multiple features. We also hoped that PCA would separate out, as orthogonal factors, aspects of prosody that truly relate to dialog from aspects with lexical, phrasal, or other significance.

4 Related Research

While dialog states have apparently not previously been tackled using PCA, other dimensionality-reduction methods have been used. Clustering has previously been applied as a way to categorize user intention-types and goals, using lexical-semantic features and neighboring-turn features as inputs (Lefevre and de Mori, 2007; Lee et al., 2009), among other methods (Gasic and Young, 2011). Hidden Markov Models have been used to identify dialog "modes" that involve common sequences of dialog-acts (Boyer et al., 2009). There is also work that uses PCA to reduce multi-factor subjective evaluations of emotion, style, or expressiveness into a few underlying dimensions, for example (Barbosa, 2009). In addition, clustering over low-level patterns of turn-taking has been used to identify a continuum of styles (Grothendieck et al., 2011). However analysis of dialog states based on prosodic features has not previously been attempted, nor has analysis of dialog behaviors over time frames shorter than the discourse or the turn sequence.

Reducing the multiplicity of prosodic features to a smaller underlying set has long been a goal for linguists. The traditional method is to start with percepts (for example, that some syllables sound louder) and then look for the acoustic-prosodic features that correlate with these perceptions. More recently the opposite tack has also been tried, starting with acoustic-prosodic features, and trying to infer a higher or deeper level of description. For example, if we discover that for many syllables pitch height, higher volume, and increased duration all correlate, then we can infer some deeper factor underlying all of these, namely stress or prominence. PCA provides a systematic way of doing this for many features at once, and it has been used for various prosodic investigations, including an exploration of the prosodic and other vocal parameters relevant to emotional dimensions (Goudbeek and Scherer, 2010) or levels of vocal effort (Charfuelan and Schröeder, 2011), categorizing glottal-flow waveforms (Pfitzinger, 2008), finding the factors involved in boundaries and accents (Batliner et al., 2001), identifying the key dimensions of variation in pitch contours using Functional Data Analysis (Gubian et al., 2010), and for purely practical purposes (Lee and Narayanan, 2005; Jurafsky et al., 2012). In our own laboratory, Justin McManus applied PCA to 4 left-context, single-speaker prosodic features, and identified the first PC with a continuum from silence to cheerful speech, and the second PC with the continuum from back-channeling to storytelling. However PCA has never before been applied to large set of features, thus we hoped it might reveal important underlying factors in prosody that have not previously been noticed: factors interactionally important, even if not salient.

5 Method

Using Switchboard, a large corpus of smalltalk between strangers over the telephone recorded in two

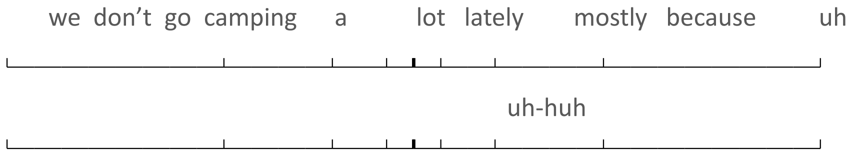

Figure 1: The 16 pitch-height feature windows, centered about a hypothetical occurrence of the word *lot* .

channels (Godfrey et al., 1992), we collected datapoints from both sides of 20 dialogs, totaling almost two hours, taking a sample every 10 milliseconds. This gave us 600,000 datapoints.

For each datapoint we computed 76 prosodic features. These features were taken from both the immediate past and the immediate future, since dialog state, by any definition, relates to both: being dependent on past context and predictive of future actions. The features were taken from both the speaker of interest and his or her interlocutor, since dialog states intrinsically involve the behavior of both parties.

Because our interest is in short-term dialog states, features were computed over only the 3-4 seconds before and after each point of interest. The sequencing of the prosodic features being obviously important, this context was split up into a sequence of windows. Wishing to give more precision and more weight to close context than more distant context, the windows closest to the point of interest were smallest, with the more distant being wider, as illustrated in Figure 1. The window sizes were fixed, not aligned with utterances, words, nor syllables.

The specific features we computed were chosen for convenience, based on a basic set previously found useful for language modeling (Ward et al., 2011). These were 1. a speaking-rate measure, over 325 millisecond windows, 2. volume, over 50 ms windows, 3. pitch height, over 150 ms windows, and 4. pitch range, over 225 ms windows. All were speaker-normalized. The values for the longer regions were obtained by simply averaging the values over two more more adjacent basic features.

In total there were 76 features: 24 volume, 20 pitch range, 16 pitch height, and 16 speaking rate. At times where there was no pitch, the average pitch value was used as substitute. All features were normalized to have mean 0 and standard deviation 1.

PCA was then done. As hoped, a few dimensions explained most of the variance, with the top 4 explaining 55%, the top 10 explaining 70%, and the top 20 explaining 81%.

We then set out to determine, for each of the dimensions, what dialog states or situations, if any, were associated with it.

Our first approach was to examine extreme datapoints. Because we thought that it would be informative to see which words tended to occur at the extremes, we filtered our datapoints to select only those which were at word onsets. For each dimension we then computed, for all of these, the values on that dimension. We then sorted these to find the highest 20 and the lowest 20. Looking at these word lists however was generally not informative, as no word or even word type predominated in any group, in fact, the words were invariably highly diverse. This perhaps indicates that the dimensions of dialog state expressed by prosody do not aligne with those expressed by words, and perhaps confirm that words can correlate with social and dialog functions in unsuspected ways (Tausczik and Pennebaker, 2010).

We next listened to some of some of these datapoints in context. First we listened to a few low-valued ones and came up with informal hypotheses about what they had in common. We then listened to more examples, winnowing and revising hypotheses as we went, until we were satisfied that we had a generalization that held for at least the majority of the cases. Then we did the same thing for the high-valued times. Finally we put the two together and found an opposition, and used this to describe the significance of the dimension as a whole. Sometimes this came easily, but sometimes it required more listening to verify or refine. This was in general easy for the top few dimensions, but more challenging for the lower ones, where the shared properties were generally weaker and more variable.

This process was unavoidably subjective, and must be considered only exploratory. We did not start out with any strong expectations, other than that many of the dimensions would relate to aspects of dialog. Our backgrounds may have predisposed us to be extra alert to turn-taking processes, but often initial hypotheses relating to turn-taking were superseded by others that explained the data better. We did not limit ourselves to terminology from any specific theoretical framework, rather we chose whichever seemed most appropriate for the phenomena.

Our second approach was to look at the loading factors, to see for each dimension which of the input prosodic features were highly correlated with it, both positively and negatively. In every case these confirmed or were compatible with our interpretations, generally revealing heavy loadings on features which previous research or simple logic suggested would relate to the dialog activities and states we had associated with the dimension.

6 Interpretations of the Top Dimensions

The results of our analyses were as follows. These must be taken as tentative, and the summary descriptions in the headings and in the tables must be read as mere mnemonics for the more complex reality that our fuller descriptions capture better, although still far from perfectly.

Dimension 1: Who's speaking?

At points with low values on this dimension the speaker of interest is speaking loudly and continuously without pause while the other is completely silent. At points with high values on this dimension the speaker of interest is producing only back-channels, while the other speaker is speaking continuously. (Points with complete silence on the part of the speaker of interest probably would have been even more extreme, but were not examined since our sample set only included timepoints where the speaker of interest was starting a word.) Unsurprisingly the features with the highest loadings were the volumes for the two speakers. Thus we identify this dimension with "who's speaking." Interestingly, of all the dimensions, this was the only with a bimodal distribution.

Dimension 2: How much involvement is there?

At points with low values on this dimension the dialog appeared to be faltering or awkward, with the lone speaker producing words slowly interspersed with non-filled pauses. High-value points were places where both speakers appeared highly involved, talking at once for several seconds, or one laughing while the other talked. Again the volume features had the highest loadings. Thus we identify this dimension with the amount of involvement.

Dimension 3: Is there a topic end?

At points with low values on this dimension there is generally a quick topic closing, in situations where the speaker had a new topic cued up and wanted to move on to it. An extreme example was when, after hearing clicks indicating call waiting, the speaker said she needed to take the other call. At points with high values on this dimension the topic was constant, sometimes with the less active participant indicating resigned boredom with a half-hearted back-channel. The features with the highest positive loadings were speaking-rate features: fast speech by the interlocutor in the near future correlated with a topic close, whereas fast speech by the current speaker about 1–2 seconds ago correlated with topic continuity. Thus we identify this dimension with topic ending.

Dimension 4: Is the referent grounded yet?

At points with low values on this dimension the speaker is often producing a content word after a filler or disfluent region, and this is soon followed by a back-channel by the other speaker. At points with high values on this dimension the speaker of interest is adding more information to make the point he wanted (starting the comment part of a topic-comment pair) sometimes after the interlocutor had responded with *oh*. Thus this dimension relates to the continuum between trying to ground something and continuing on with something already grounded. Trying to ground correlated with an upcoming fast speaking rate, while proceeding after grounding correlated with a high volume. Thus we identify this dimension with the degree of grounding.

Dimension 5: Does the speaker want to start or stop?

At points with low values on this dimension the speaker of interest is starting a turn strongly, sometimes as a turn-grab or even cutting-off the other speaker. At points with high values on this dimension the speaker is strongly yielding the turn, coupled with the interlocutor very swiftly taking up the turn. Often the turn yield occurs when the speaker is soliciting a response, either explicitly or by expressing an opinion that seems intended to invoke a response. As might be expected, cut-offs correlate with high volume on the part of the interrupting speaker, while clear turn yields correlate with past high volume on the part of the speaker who is ending. Thus we identify this dimension with starting versus stopping.

Dimension 6: Has empathy been expressed yet?

At points with low values on this dimension the speaker is continuing shortly after a high-content, emotionally-colored word that has just been acknowledged by the interlocutor. At points with high values on this dimension, the speaker is acknowledging a feeling or attitude just expressed by the other, by expressing agreement with a short turn such as *that's right* or *yeah, Arizona's beautiful!*. Continuing after empathic grounding correlated with high volume after a couple of seconds; expressing empathy with a short comment correlated with the interlocutor recently having produced a word with high pitch. Thus we identify this dimension with the degree of empathy established.

Dimension 7: Are the speakers synchronized?

At points with low values on this dimension both speakers inadvertently start speaking at the same time. At points with high values on this dimension the speakers swiftly and successfully interleave their speaking, for example by completing each other's turns or with back-channels. The features with the highest positive loadings were those of pitch range and speaking rate with the volume factors having mostly negative loadings. Thus we identify this dimension with the degree of turn synchronization.

Dimension 8: Is the turn end unambiguous?

At points with low values on this dimension the speaker is dragging out a turn which appears, content-wise, to be already finished, producing post-completions, such as *uh* or *or anything like that*. At points with high values on this dimension, often the speaker is definitively ending a turn. The feature with the highest positive loading was pitch range, unsurprisingly since clear turn ends often involve a sharp pitch fall. Thus we identify this dimension with the degree of ambiguity of the turn end.

Dimension 9: Is the topic exhausted?

At points with low values on this dimension a speaker is closing out a topic due to running out of things to say. Often at points with high values on this dimension the speaker is staying with one topic, with continuing interest also from the interlocutor. The most positively correlated feature was the interlocutor's volume 400–800 ms ago, for example during a back-channel or comment showing interest. Thus we identify this dimension with the degree of interest in the current topic.

Dimension 10: Is the speaker thinking?

At points with low values on this dimension the speaker is looking for a word, choosing her words carefully, or recalling something, typically inside a turn but preceded by a short pause or an *um*. At points with high values on this dimension the speaker seems to be giving up on the topic, declaiming any relevant knowledge and/or yielding the turn. The features correlating most with the memory-search/lexical-access state were those of high volume by the speaker 50–1500 milliseconds later; the features correlating most with the giving-up state were speaking rate. Thus we identify this dimension with the degree to which the speaker is putting mental effort into continuing.

Dimension 11: How quick-thinking is the speaker?

Points with low values on this dimension included two types: first where a speaker is ending a false start and about to start over, and second where the speaker is about to be cut off by the interlocutor while saying something noncommittal to end a turn, such as *I guess*. Points with high values included swift echos and confirmations, which seemed to reflect quickness and dominance. Thus we identify this dimension with quickness, confidence and dominance versus the lack thereof.

Dimension 12: Is the speaker claiming or yielding the floor?

Points 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. Slow future speaking rate, by both speakers, aligned with the low values, and fast rate with the high values. We identify this dimension with the floor claim/yield continuum.

Dimension 13: How compatible is the proposition with the context?

Points with low values on this dimension occurred in the course of a self-narrative at the beginning of something contradicting what the listener may have inferred, or actually did think and say, for example with *no, we actually don't*. Points with high values of this dimension generally involved a restatement of something said before either by the speaker or the interloctor, for example restating a question after the other failed to answer, or opining that a football team can now expect a few bad years, just a dozen seconds after the interlocutor had already expressed essentially the same thought. The low, contradicting side had high volume and slow speaking rate for a fraction of a second; the restatements were the opposite. Thus we identify this dimension with the continuum between a contrast-type rhetorical structure and a repetition-type one.

Dimension 14: Are the words being said important?

Points with low values on this dimension occur when the speaker is rambling: speaking with frequent minor disfluencies while droning on about something that he seems to have little interested in, in part because the other person seems to have nothing better to do than listen. Points with high values on this dimension occur with emphasis and seemed bright in tone. Slow speaking rate correlated highest with the rambling, boring side of the dimension, and future interlocutor pitch height with the emphasizing side. Thus we identify this dimension with the importance of the current word or words, and the degree of mutual engagement.

Dimension 15: Are the words premature or delayed?

Points with low values on this dimension included examples where the speaker is strongly holding the floor despite a momentary disfluency, for example *uh and* or *well it's it's difficult*, using creaky voice and projecting authority. Points with high value on this dimension overlapped substantially with those high on dimension 14, but in addition seemed to come when the speaker starts sharing some information he had been wanting to talk about but saving up, for in a drawn-out political discussion, a new piece of evidence supporting an opinion expressed much earlier. Thus we identify this dimension with the continuum between talking as soon as you have something to say (or even slightly before) versus talking about something when the time is ripe.

Dimension 16: How positive is the speaker's stance?

Points with low values on this dimension were on words spoken while laughing or near such words, in the course of self-narrative while recounting a humorous episode. Points with high values on this dimension also sometimes occurred in a self narratives, but with negative affect, as in *brakes were starting to fail*, or in deploring statements such as *subject them to discriminatory practices*. Low values correlated with a slow speaking rate; high values with the pitch height. This we identify this a humorous/regrettable continuum.

Other Dimensions

Space does not permit the discussion of further dimensions here, but the end of Table 1 and Table 2 summarize what we have seen in some other dimensions that we have examined for various reasons, some discussed elsewhere (dimensions 25, 62, and 72 in (Ward and Vega, 2012 submitted) and 17, 18, 21, 24, 26, and 72 in (Ward et al., 2012 submitted)). Of course, not all dimensions are mostly about dialog, for example dimension 29 appears to be described best as relating simply to the presence or absence of a stressed word (Ward et al., 2012 submitted), although that of course is not without implications for what dialog activities may cooccur.

7 Discussion

Although prosody is messy and multifunctional, this exploration shows that PCA can derive from raw features a set of dimensions which explain much of the data, and which are surprisingly interpretable.

1	this speaker talking vs. other speaker talking	32%
2	neither speaking vs. both speaking	9%
3	topic closing vs. topic continuation	8%
4	grounding vs. grounded	6%
5	turn grab vs. turn yield	3%
6	seeking empathy vs. expressing empathy	3%
7	floor conflict vs. floor sharing	3%
8	dragging out a turn vs. ending confidently and crisply	3%
9	topic exhaustion vs. topic interest	2%
10	lexical access or memory retrieval vs. disengaging	2%
11	low content and low confidence vs. quickness	1%
12	claiming the floor vs. releasing the floor	1%
13	starting a contrasting statement vs. starting a restatement	1%
14	rambling vs. placing emphasis	1%
15	speaking before ready vs. presenting held-back information	1%
16	humorous vs. regrettable	1%
17	new perspective vs. elaborating current feeling	1%
18	seeking sympathy vs. expressing sympathy	1%
19	solicitous vs. controlling	1%
20	calm emphasis vs. provocativeness	1%

Table 1: Interpretations of top 20 dimensions, with the variance explained by each

21	mitigating a potential face threat vs. agreeing, with humor
24	agreeing and preparing to move on vs. jointly focusing
25	personal experience vs. second-hand opinion
26	signalling interestingness vs. downplaying things
62	explaining/excusing oneself vs. blaming someone/something
72	speaking awkwardly vs. speaking with a nicely cadenced delivery

Table 2: Interpretations of some other dimensions

Overall, the top dimensions covered a broad sampling of the topics generally considered important in dialog research. This can be taken to indicate that the field of dialog studies is mostly already working on the important things after all. However previously unremarked aspects of dialog behavior do appear to surface in some of the lower dimensions; here further examination is needed.

We had hoped that PCA would separate out the dialog-relevant aspects of prosody from the aspects of prosody serving other functions. Generally this was true, although in part because the non-dialog functions of prosody didn't show up strongly at all. While this was probably due in part to the specific feature set used, it still suggests that dialog factors are overwhelmingly important for prosody. Partial exceptions were emotion, attitude, rhetorical structure, speaking styles and interaction styles, all of which appeared as aspects of some dimensions. Some dimensions also seemed to relate to dialects, personality traits, or individuals; for example, many of the most unambiguous turn endings (dimension 8) were by the same few speakers, who seemed to us to be businesslike and dominant.

8 Potential Applications

These dimensions, and similar empirically-derived sets, are potentially useful for various applications.

First, the inferred dimensions could serve as a first-pass specification of the skills needed for a competent dialog agent: suggesting a dialog manager whose core function is to monitor, predict, and guide the development of the dialog in terms of the top 10 or so dimensions. This technique could be very generally useful: since it supports the discovery of dialog dimensions in a purely data-driven way (apart from the subjective interpretations, which are not always needed), this may lead to methods for the automatically generation of dialog models and dialog managers for arbitrary new domains.

Second, for generation and synthesis, given the increased interest in going beyond intelligibility to also give utterances dialog-appropriate wordings and realizations, the inferred dimensions suggest what is needed for dialog applications: we may have identified the most important parameters for adapting and controlling a speech synthesizer's prosodic behavior for dialog applications.

Third, dimensional representations of dialog state could be useful for predicting the speaker's upcoming word choices, that is, useful for language modeling and thus speech recognition, as an improvement on dialog-act descriptions of state or descriptions in terms of raw, non-independent prosodic features (Shriberg and Stolcke, 2004; Ward et al., 2011; Stoyanchev and Stent, 2012). Initial results of conditioning on 25 dimensions gave a 26.8% perplexity reduction (Ward and Vega, 2012 submitted).

These dimensions could also be used for other purposes, including a more-like-this function for audio search based on similarity in terms of dialog context; better characterizing the functions of discourse markers; tracking the time course of action sequences leading to impressions of dominance, friendliness and the like; finding salient or significant events in meeting recordings; and teaching second language learners the prosodic patterns of dialog.

9 Future Work

Our study was exploratory, and there are many obvious ways to improve on it. It would be good to apply this method using richer feature sets, including for example voicing fraction, pitch slope, pitch contour features, spectral tilt, voicing properties, and syllable- and word-aligned features, to get a more complete view of what prosody contributes to dialog. Going further, one might also use temporal features (Ward et al., 2011), features of gaze, gesture, and words, perhaps in a suitable vector-space representation (Bengio et al., 2003). Better feature weighting could also be useful for refining the ranking of the dimensions: while our method treated one standard deviation of variance in one feature as equal in importance to one standard deviation in any other, in human perception this is certainly not the case. It would also be interesting to apply this method to other corpora in other domains: for example in task-oriented dialogs we might expect it to find additional important dimensions relating to task structure, question type, recovery from misunderstandings, uncertainty, and so on. Finally, it would be interesting to explore which of these dimensions of state actually matter most for dialog success (Tetreault and Litman, 2006).

In addition to the identification of specific dimensions of dialog in casual conversations, this paper contributes a new method: that of using PCA over low-level, observable features to identify important dimensions of dialog state, which could be applied more generally.

While we see numerous advantages for quantitative, dimensional dialog state modeling, we do not think that this obsoletes more classical methods. Indeed, it would be interesting to explore how commonly used dialog states and acts relate to these dimensions; for example, to take the set of utterances labeled wh-questions in NXT Switchboard and examine where they are located in the "dialog space" defined by these dimensions (Calhoun et al., 2010; Ward et al., 2012 submitted).

Acknowledgments

This work was supported in part by NSF Award IIS-0914868. We thank Olac Fuentes for suggesting PCA, Justin McManus for the prototype analysis, Shreyas Karkhedkar for help with the basic features, and David Novick for discussion.

References

Plinio Barbosa. 2009. Detecting changes in speech expressiveness in participants of a radio program. In *Interspeech*, pages 2155–2158.

Anton Batliner, Jan Buckow, Richard Huber, Volker Warnke, Elmar Nöth, and Heinrich Niemann. 2001. Boiling down prosody for the classification of boundaries and accents in German and English. In *Eurospeech*, pages 2781–2784.

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Kristy Elizabeth Boyer, Eun Young Ha, Robert Phillips, Michael D. Wallis, Mladen A. Vouk, and James C. Lester. 2009. Inferring tutorial dialogue structure with hidden Markov modeling. In *Proc. NAACL-HLT Workshop on Innovative Uses of NLP for Building Educational Applications*, pages 19–26.

Harry Bunt. 2011. Multifunctionality in dialogue. *Computer Speech and Language*, 25:222–245.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, et al. 2010. The NXT-format Switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44:387–419.

Marcela Charfuelan and Marc Schröeder. 2011. Investigating the prosody and voice quality of social signals in scenario meetings. In *Proc. Affective Computing and Intelligent Interaction*.

Milica Gasic and Steve Young. 2011. Effective handling of dialogue state in the hidden information state POMDP-based dialogue manager. *ACM Transactions on Speech and Language Processing*, 7.

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

Martijn Goudbeek and Klaus Scherer. 2010. Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *Journal of the Acoustical Society of America*, 128:1322–1336.

John Grothendieck, Allen L. Gorin, and Nash M. Borges. 2011. Social correlates of turn-taking style. *Computer Speech and Language*, 25:789–801.

Michelle Gubian, Francesco Cangemi, and Lou Boves. 2010. Automatic and data driven pitch contour manipulation with functional data analysis. In *Speech Prosody*.

Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. 2012. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech and Language*, in press.

Chul Min Lee and Shrikanth Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13:293–303.

Cheongjae Lee, Sangkeun Jung, Kyungduk Kim, and Gary Geunbae Lee. 2009. Automatic agenda graph construction from human-human dialogs using clustering method. In *Proc. NAACL-HLT 2009: Short Papers*, pages 89–92.

Fabrice Lefevre and Renato de Mori. 2007. Unsupervised state clustering for stochastic dialog management. In *ASRU*, pages 550–553.

Scott McGlashan, Daniel C. Burnett, et al. 2010. Voice extensible markup language (VoiceXML) 3.0. Technical report, W3C.

Hartmut R. Pfitzinger. 2008. Segmental effects on the prosody of voice quality. In *Acoustics'08*, pages 3159–3164.

Elizabeth Shriberg and Andreas Stolcke. 2004. 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*, pages 105–114. Springer-Verlag.

Svetlana Stoyanchev and Amanda Stent. 2012. Concept type prediction and responsive adaptation in a dialogue system. *Dialogue and Discourse*, 3.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24–54.

Joel R. Tetreault and Diane J. Litman. 2006. Comparing the utility of state features in spoken dialogue using reinforcement learning. In *HLT-NAACL*, pages 272–279.

David Traum and S. Larsson. 2003. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie Smith, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer.

Nigel G. Ward and Alejandro Vega. 2012, submitted. Towards empirical dialog-state modeling and its use in language modeling. In *Interspeech*.

Nigel G. Ward, Alejandro Vega, and Timo Baumann. 2011. Prosodic and temporal features for language modeling for dialog. *Speech Communication*, 54:161–174.

Nigel G. Ward, David G. Novick, and Alejandro Vega. 2012, submitted. Where in dialog space does uh-huh occur? In *Interdisciplinary Workshop on Feedback Behaviors in Dialog at Interspeech 2012*.