Prosodic and Temporal Features for Language Modeling for Dialog

Nigel G. Ward[a], Alejandro Vega[a], Timo Baumann[b]

[a]*Computer Science, University of Texas at El Paso*
500 West University Avenue, El Paso, Texas 79968 USA
[b]*University of Potsdam, Linguistics Department*
Karl-Liebknecht-Straße 24, 14476 Potsdam, Germany.

Abstract

If we can model the cognitive and communicative processes underlying speech, we should be able to better predict what a speaker will do. With this idea as inspiration, we examine a number of prosodic and timing features as potential sources of information on what words the speaker is likely to say next. In spontaneous dialog we find that word probabilities do vary with such featues. Using perplexity as the metric, the most informative of these included recent speaking rate, volume, pitch height and range, time since start of utterance, and time since end of the interlocutor's last utterance. Using simple combinations of such features to augment trigram language models modestly reduced perplexities on both Switchboard and Verbmobil II.

Key words: dialog dynamics, dialog state, prosody, interlocutor behavior, word probabilities, prediction, perplexity, speech recognition, Switchboard corpus, Verbmobil corpus

1. Introduction

In interpersonal dynamics, the human ability to predict the micro-level, moment-by-moment, actions of an interlocutor has been identified as a central issue in coordination (Sebanz et al., 2006; Barsalou et al., 2007; Streek and Jordan, 2009), and better predictions have been seen to correlate with more empathy and success in interactions (Gratch et al.,

☆We thank Shreyas Karkhedkar, Nisha Kiran, Shubhra Datta and Gary Beverungen for contributing to the code and analysis; David Novick, Olac Fuentes and several Interspeech, ASRU, Speech Communication and National Science Foundation reviewers for comments; and Joe Picone and Andreas Stolcke for making available the Switchboard labeling and the SRILM toolkit, respectively. This work was supported in part by NSF Grants IIS-0415150 and IIS-0914868 and REU supplements thereto, and by the US Army Research, Development and Engineering Command, via a subcontract to the USC Institute for Creative Technologies.

Email addresses: nigelward@acm.org (Nigel G. Ward), avega5@miners.utep.edu (Alejandro Vega), timo@ling.uni-potsdam.de (Timo Baumann)

URL: www.nigelward.com (Nigel G. Ward)

2006; Jahr and Eldevik, 2007; Beebe et al., 2008; Macrae et al., 2008). Language modeling can be seen as such a prediction problem, where we need to predict the speaker's next word, given the previous words and other prior context. Having good language models is important, not least because every speech recognizer relies on one to provide estimates of the probabilities of the word hypotheses it searches through.

In the classical formulation, the task of a language model is "to compute, for every word string, W, the *a priori* probability P(W)" (Jelinek, 1997). This statement embodies the assumption that only lexical context matters, with other information, such as durations, timing, pitch, and detailed phonetics, being seen as relevant only to the acoustic model. Today most language models make this assumption, treating speech as simply sequences of words, but for spontaneous speech and dialog it is an oversimplification (Ji and Bilmes, 2010). It is probably not coincidental that speech recognizer performance is still weak for spontaneous speech in general, and dialog in particular; and there is evidence from a human-subjects experiment that languages models for dialog can be improved more by using additional sources of information than by just improving the modeling of lexical context (Ward and Walker, 2009).

Fujisaki (2008) has argued that accurately recognizing speech ultimately requires "mind modeling." While this is a very long-term goal, we may still reap short-term benefit from models inspired by the fact that spoken language is created by human minds and for human minds. From this perspective, this paper explores the value of new sources of information for language modeling. Section 2 explains why we expect information about the speaker's cognitive and communicative states, moment-by-moment, to be revealed by non-lexical information. Section 3 presents our model for representing the way word probabilities depend on previous non-lexical context and shows how this information can be combined with that given by a standard n-gram model. Section 4 examines 23 non-lexical features, related to timing and prosody, and shows that all but one provide useful information, as measured by perplexity reductions. Section 5 shows how a simple combination of such features can be used in a recognizer. Finally, Sections 6 summarizes and notes directions for future work.

2. Time, States, and Events in Dialog

Our modeling strategy is inspired by a theme central to many recent psycholinguistic studies of spoken dialog: that it is a process in time (Clark, 1996, 2002). However this aspect still often escapes attention, perhaps in part because so much research relies on written representations which abstract away from time. Consider for example four utterances, first as word sequences: A: *"they th- they a- after five o'clock they uh the the uh daycare workers are pretty burned out,"* B: *"yeah,"* A: *"and so they they wheel out the T.V. and put the kids in front of the T.V.,"* B: *laugh*; and then in a richer notation showing the times of occurrence, in Figure 1. Looking at a dialog in this way, the temporal properties, especially the variation in word lengths and pause lengths, pop out as potentially significant.

Some of the cognitive and communicative processes that underlie dialog seem evident even in this little example. The degree of fluency varies, with apparent spurts of fluency interleaved with fillers, lengthened words and silences, presumably reflecting underlying

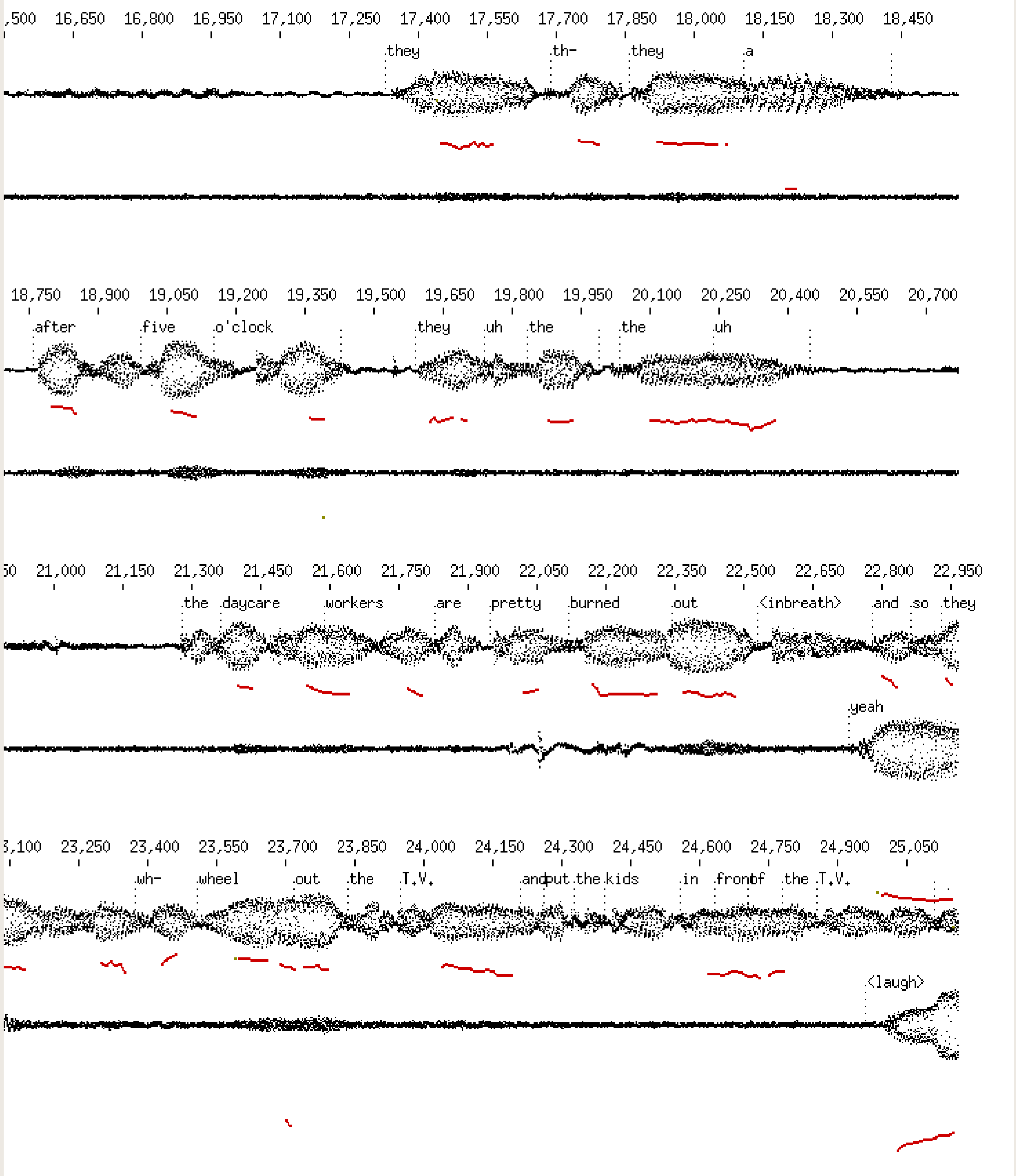

Figure 1: Conversation Fragment. Each of the four strips includes a timeline and two rows, one per speaker. Each row includes a transcription, the signal and the pitch. The second speaker's pitch range is so low that his pitch contours appear far below his signal. This fragment occurred after some talk about television-watching habits and effects on children. Audio for this clip is available at http://cs.utep.edu/nigel/abstracts/prosody-lm.html.

processes of deciding what to say and of formulating it. There is also turn management: the primary speaker appears to be signaling his intention to hold the floor, with occasional invitations to the other to back-channel or otherwise respond. In other dialogs there is also variation over time in the speaker's degree of involvement, of valence (positive or negative attitude), and of dominance, to name just three factors. There are also of course reflections of syntactic, semantic, and discourse-structuring processes.

From the state at any given time, it seems that it should be possible to predict, to some extent, what words are likely to occur. For example, at 24 seconds into this dialog, it seems that this speaker has attained momentary fluency (with the past few words being pronounced without problem and as part of a well-formed intonation contour), so the next few words are unlikely to be fillers or disfluency markers. It seems that he's been speaking for a while with only a perfunctory contribution from his interlocutor (the *yeah*), so the next few words are likely to include affective or evaluative words, or perhaps a turn yield. We envisage a system which is able to model, moment-by-moment, the state of a speaker, and from that, to be able to predict the upcoming words, as illustrated by Figure 2. If we can accomplish this, the resulting language model could turn out to be more robust than n-grams, which are known to be brittle (Bellegarda, 2004); to the extent that we can represent patterns of word occurrence that reflect fundamental cognitive processes and constraints, such language models may transfer well across domains and tasks.

Our notion of how to gain leverage for language modeling can thus be summarized in four principles:

1. the state of the speaker varies over time,
2. the state is complex,
3. the likely state can be inferred in part from non-lexical context, and
4. this state is somewhat predictive of what word the speaker will say next.

These principles guide the exploration of features and the construction of models in this paper.

2.1. Research with Similar Motivations

Other researchers have previously approached the language modeling problem with an interest in the effects of speaker states and in using information beyond the word sequence. This subsection discusses three landmark studies.

Stolcke, Shriberg and colleagues (Stolcke et al., 1999) modelled the "hidden events" in speech, that is, events which are not explicit in the word sequence, but are nevertheless significant, both to the speaker and the hearer. In particular they used sentence boundaries and disfluency points. They showed how, for language modeling purposes, these hidden events could be modeled as additional "words" in the word sequence. Since this formulation fits easily into the n-gram framework, future word probabilities could be conditioned on these events. The hidden events were detected using both lexical and prosodic context. Incorporated into a speech recognizer for the Switchboard corpus (Godfrey et al., 1992), the model gave a 0.9% absolute (2% relative) decrease in word error rate.

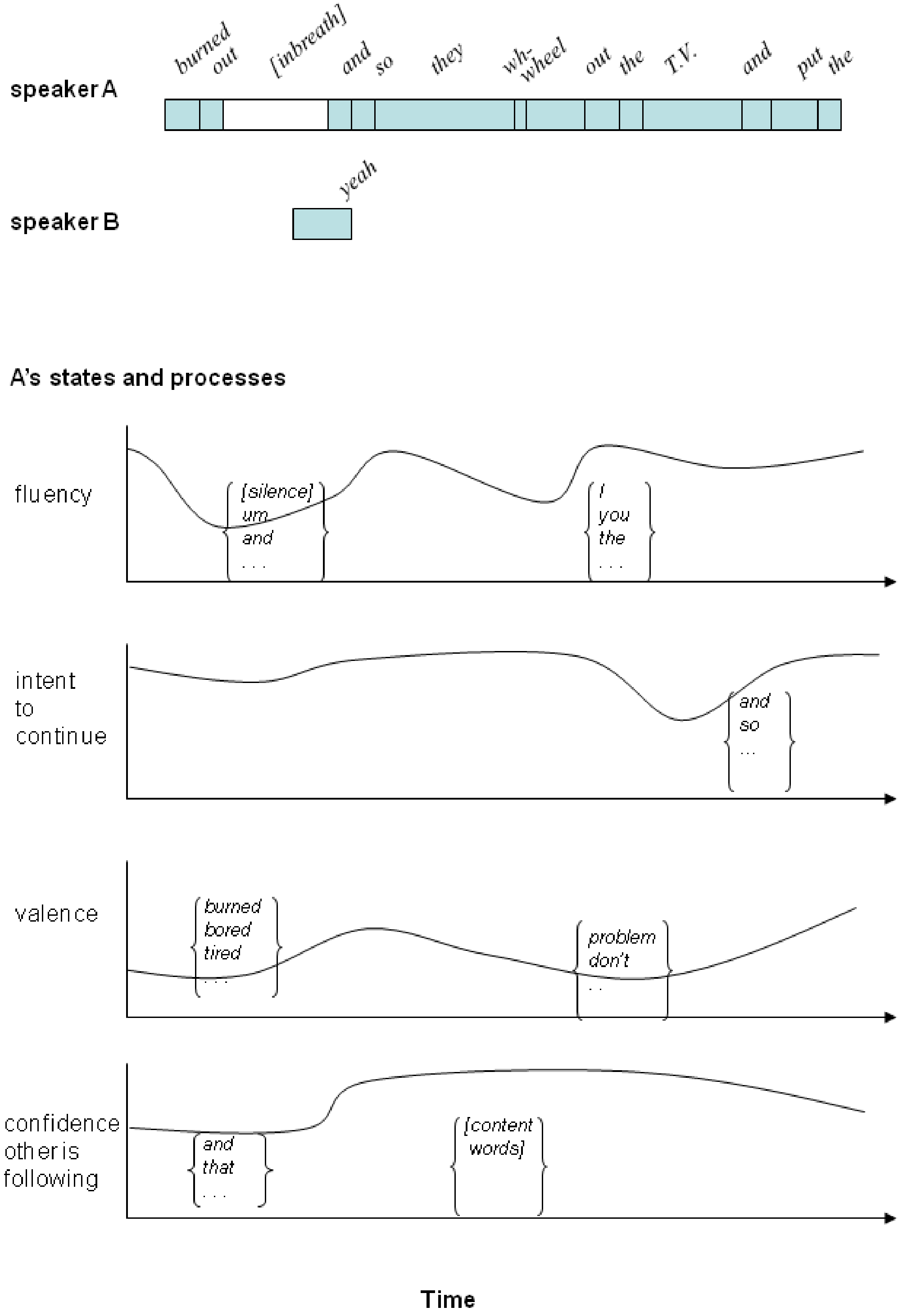

Figure 2: A fanciful image of the aspects of the cognitive state of the speaker in Figure 1 from 22 to 24.5 seconds. Each curve indicates the varying intensity over time of a cognitive state, process, or need. The words in brackets suggest some of words more likely to be produced while in that state.

Although impressive, the model has limitations which make it hard to generalize. The hidden event formulation is suitable only for modeling underlying cognitive events which are binary, either present or absent, rather than graded; which occur at a single instant, rather than lasting over some time interval or attenuating or growing over time; and which are mutually exclusive, rather than being co-present with other factors. Regarding the last point, there is ample evidence for the presence of multiple dimensions of emotional, attitudinal, meta-communication and interpersonal communication in parallel to the communication of content (Goffman, 1981; Clark, 1996; Brennan and Hulteen, 1995; Campbell, 2007; Petukhova and Bunt, 2009), and evidence that the processes of formulating and speaking and the processes of hearing and comprehending, although largely temporally separate and distinct, can operate in parallel in certain limited ways (Yngve, 1970; Jaffe, 1978; Bard et al., 2002). In short, the hidden-event model does not support our principles 1 and 2 above.

Ma and colleagues (Ma et al., 2000), noting that one can build language models specific for certain domains or for certain dialog acts, proposed an even finer-grained decomposition, breaking each utterance into given and new parts, and using a separate language model for each. This technique gave a 0.3% decrease in word error rate. In a sense, our approach here is to experiment with even finer-grained decompositions.

Ji and Bilmes noted that the behavior of a speaker can be affected by the recent behavior of the interlocutor. In particular, they observed that the immediately preceding word of the interlocutor can help predict the next word of the speaker, mostly, it seems, due to conversational routines and semantic priming effects, and that modeling this gives up to a 8.9% perplexity reduction on Switchboard (Ji and Bilmes, 2004). They also developed a way to use information about the dialog act of the interlocutor's previous utterance, and obtained a 0.5% (2% relative) reduction in word error rate (Ji and Bilmes, 2010).

Thus we are not alone wanting language models that consider more than just word sequence information. In this paper we seek to generalize the results of previous research, in line with the principles proposed above. Our approach is exploratory, and our methods are accordingly novel in two main ways. First, we choose not to use theoretical or *a priori* constructs, such as disfluency point, given information, or dialog act. Although demonstrably useful, such constructs may ultimately be limiting. Instead, we directly condition the word probabilities on the observables, doing without mediating variables. Second, we chose to start with an open search for correlations and patterns, initially unconstrained by considerations of what would be computationally convenient for extending existing language models.

3. Methods

This section presents our methods for discovering, modeling, and applying regularities in the way that word probabilities vary with non-lexical features. The running example will be the feature "time into utterance."

0.0–0.5s	0.5–1.0s	1.0–1.5s	1.5–2.0s	2.0–2.5s	. . .	4.0–4.5s	. . .	8.0–8.5s	. . .	16.0-16.5s
yeah	I	I	I	and		and		and		and
I	the	the	the	the		I		I		the
and	a	and	and	I		the		the		I
you	to	a	to	to		a		to		you
uh	and	to	a	a		to		you		to

Table 1: The Five Most Frequent Words in Selected Buckets. Times are in seconds.

Figure 3: Ratio of Frequency in Specific Time Buckets to Overall Frequency (R) vs. Time Into Utterance for the Three Most Common Words. Words at utterance start (time into utterance = 0.0) are excluded. The rightmost points represent the range 9.5 seconds and up.

3.1. Time into Utterance: Initial Observations

Over the course of a typical utterance it seems likely that a speaker will generally go through various states — including turn grabbing, referring to given information, presenting new information, assessing or expressing an attitude about the new information, and yielding the turn, possibly interleaved with disfluent interludes — that these states will relate to time into utterance, and that these states will affect the words spoken.

To investigate this we used (here and throughout this paper) the Switchboard corpus, a collection of short telephone conversations on light topics between mostly unacquainted adults, with the ISIP transcriptions, which are time-aligned at the word level (Godfrey et al., 1992; ISIP, 2003). We split each track into utterances, initially defined as sequences of words delimited by at least 1 second of silence both before and after, using the regions labeled *[silence]* in the transcripts and merging adjacent silence regions.

For each word we marked the time from the start of utterance to the start of the word. Conceptually each utterance was split into buckets. For example, words that began between 0 and 0.1 seconds into the utterance were counted as belonging to bucket 0, those between 0.1 and 0.2 seconds as belonging to bucket 1, etc. We computed the probability of each word in each bucket, the "bucket probability" (time-based probability) $P_{tb}(w_i@t)$ for each word, as its count in the bucket for t divided by the total in that bucket:

$$P_{tb}(w_i@t) = \frac{count(w_i@t)}{\sum_j count(w_j@t)} \tag{1}$$

Table 1 shows that the most common words do indeed vary with time into utterance. To more clearly see the tendencies of words to appear in different buckets, we then computed the ratio of this time-based probability to the standard unigram probability:

$$R(w_i@t) = \frac{P_{tb}(w_i@t)}{P_{unigram}(w_i)} \tag{2}$$

Figure 3 illustrates how this ratio can vary over time.

3.2. Meaningful States and Shallow States

As hoped, it seems that some of these probability variations are reflections of cognitive process constraints. For example, the fact that low frequency words, typically content-rich words, are relatively more common later in utterances, may be because they are harder to retrieve from the mental lexicon or because they are easier for listeners to process if heard later in an utterance. The fact that the word *know* grows in frequency over time, being less than 1.4% over the first 5 seconds but over 1.8% after 10 seconds, is perhaps because of the time it takes to reason about knowledge states. The distribution of *think* is quite different: its likelihood is high early in utterances but drops over time. The distribution of the word *I* is also interesting, although harder to explain with confidence: it occurs twice as often near the start of utterances as elsewhere, peaking at around 0.2 seconds in, and *I* is far more common initially than *you*, but the difference narrows over time, perhaps because it is easier to talk about oneself, as doing so usually requires no inference, only retrieval.

But there are also probability variations that do not relate to any cognitive process or facts known to us, such as the fact that times and dates are relatively common only after about 2 seconds in (Table 2).

Fortunately, making better predictions does not require an understanding of the underlying cognitive dynamics: we can directly use the probability patterns instead. That is, we can simply compute the probability for each word in each bucket, and then use that in a language model, as detailed in the next section. This represents a strategic retreat from the goal of developing true cognitive models, but avoids many difficult problems, including those of defining, identifying, delimiting, and hand-labeling or inferring the underlying states. Thus the states examined in this paper are all shallow ones, defined in terms of objective, observable events. This makes them easy to compute from the data and this enables many new sorts of regularities to be represented, including some, such as "number words are relatively common 3 to 10 seconds into an utterance," that could not be handled using previous methods.

3.3. Combination with N-grams

To determine the utility of time into utterance, or any new feature, for language modeling, we are less interested in its information content in isolation than in whether it proves information which is usefully complementary to that provided by existing methods.

This subsection presents a way to use the information provided by a new feature to augment an existing language model, to determine what additional value it may bring. While this method could be used with any language model, we illustrate its use with a trigram model, specifically the SRILM implementation including backoff.

Our first attempt to use time-into-utterance information combined it with the trigram model by linear interpolation, for each word generating a probability estimate using a simple weighted average of the bucket probability (Equation 1) and the trigram probability. However, this performed poorly; as the trigram probability estimates were generally quite good, crudely averaging them with a weaker model was counterproductive.

We therefore decided to use the time-based probabilities merely to tweak the trigram probabilities, using a scaling factor derived from R to determine how much to tweak. For

bucket	high-S words	low-S words
ϵ–0.1s	don't, that's, know, think, well, you, was, yeah, do, have …	with, out, or, on, be
0.1–0.2s	okay, yes, that's, sure, yeah, really, haven't, don't, think, can't …	over, money, every, care, anything
0.2–0.3s	okay, right, yes, yeah, no, see, that's, sure, i-, well …	minutes, [laughter-okay], home, everything, dear
0.3–0.4s	great, right, uh-huh, yes, well, okay, no, that's, haven't, yeah …	day, school, things, year, bit
0.4–0.5s	uh-huh, great, right, okay, well, yes, yeah, that's, no, good …	come, stuff, her, every, day
0.5–1.0s	um-hum, uh-huh, agree, yeah, huh, yes, definitely, okay, heard, well …	jury, child, whether, weeks
1.0–1.5s	uh-huh, huh, bye-bye, bet, um-hum, yeah, exactly, isn't well, oh …	involved, education, during
1.5–2.0s	bye-bye, huh, friends, talked, yes, problem, funny, age, tell …	education, change, twelve, hand
2.0–2.5s	today, night, huh, though, mine, supposed, while, Texas, remember …	places, might, couldn't, moved
2.5–3.0s	Texas, times, program, huh, high, movie, insurance, system, enjoy …	feel, life, best, whatever, stay
3.0–3.5s	until, college, usually, basically, ago, try, gone, lived, made, fact …	person, percent, thinking, thirty
3.5–4.0s	thirty, myself, huh, week, part, lived, last, state, spend, run …	um-hum, fun, thinking, great, enjoy
4.0–4.5s	call, month, took, usually, movie, called, child, Texas, ten, someone …	being, um-hum, own, goes, huh
4.5–5.0s	movie, since, system, started, life, working, might, point, doing …	great, um-hum, may, love, am
5.0–5.5s	couple, college, years, times, bit, whatever, money, year, both, Dallas …	okay, still, gets, away, idea
5.5–6.0s	ago, somebody, times, year, try, college, actually, least, I'll, being …	great, okay, may, interesting, love
6.0–6.5s	country, own, does, while, pay, need, everything, husband, went, stuff …	started, great, anyway, yes
6.5–7.0s	few, look, house, care, away, why, watch, hundred, couple, enough …	sometimes, um-hum, started
7.0–7.6s	ago, week, has, always, being, whatever, try, times, six, wasn't …	area, oh, also, yes, uh-huh
7.5–8.0s	four, wasn't, usually, different, better, take, most, few, after, two …	um-hum, yes, another, uh-huh
8.0–8.5s	whatever, everything, having, through, being, come, stuff, first, either …	too, did, uh-huh, um-hum
8.5–9.0s	dollars, come, were, house, five, twenty, these, last, first, before …	okay, oh, live, interesting, um-hum
9.0–9.5s	his, hard, these, different, doesn't, sort, before, back, school, live …	right, yeah, okay, will, um-hum
9.5s-∞	authority, shirts, obvious, whereas, pants, corn, losing, bottle, percentage …	[laughter-okay], dear

Table 2: Characteristic and Uncharacteristic Words in Various Time-into-Utterance Buckets, that is, words with the highest and lowest scaling factors S.

example, for a word occurring at time t, if the bucket probability R (Equation 1) indicates that the word is more common at t than at other times, then we multiply the trigram probability by a scaling factor S to reflect this. This gives the "bucket-scaled" trigram probabilities:

$$P_{bs}(w_i@t|c) = S(w_i@t)P_{trigram}(w_i|c) \tag{3}$$

where c is the local context, for trigrams specifically just the preceding two words.

The scaling factor is based on R indirectly rather than directly, for two reasons. First, R is less informative in cases where the bucket probability is based on sparse counts, as for infrequent words or words in late buckets. To estimate the informativeness we use the χ^2 test to evaluate the hypothesis that the number of occurrences of the word in the bucket differs from that expected, which is just the product of the bucket size and the word's unigram probability. We compute the P-value of this hypothesis, p, and from that our confidence in the hypothesis: $q = 1 - p$. (If the expected count of the word in the bucket is less than 5, then we have no confidence, and set q to 0.) We then use this confidence measure q to derive the scaling factor S: it depends on R to the q^{th} power. Thus, if the confidence in the bucket probability is low, then S will be close to 1 and the time-based information will have little effect. (In particular, if there are less than 5 occurrences of a word in a bucket, as is almost always the case, the tweaking is effectively a null operation.)

The second complication in the computation of S arises from the fact that the time-based estimate and the trigram estimate are not independent. Even if we are confident that the bucket-based probability for some word is a better estimate than the unigram probability, that does not imply that it is also a better estimate than the trigram probability. It is therefore necessary to reduce the weight of the bucket-based probability relative to the trigram probabilities. This is done by raising R to a constant power k less than 1, where a suitable value for k is determined empirically.

$$S(w_i@t) = R(w_i@t)^{kq} \tag{4}$$

One more necessary detail is smoothing. While proper smoothing is important for good performance, here we do the simplest thing possible: if the count in some bucket for some word is 0 we replace it with 1. This ensures that R is never 0, which is required to make S in equation 4 always tend to 1 as the time-based information gets weaker. No explicit discounting is done, since discounting happens as a side-effect of normalization. Table 2 shows words with extreme S values in each bucket.

Finally there is a normalization step to ensure that all the probabilities across the vocabulary add to 1 in each case, that is, for every combination of lexical context and bucket. This is done at runtime: when looking up the probability for a word, the bucket-scaled trigram probabilities for all the words in the corpus are computed, and the bucket-scaled trigram probability of the word of interest is divided by the sum. This gives the normalized combined probability, P_n:

$$P_n(w_i@t|c) = \frac{P_{bs}(w_i@t|c)}{\sum_j P_{bs}(w_j@t|c)} \quad (5)$$

Since the values of P_{bs} depend on the preceding words as well as the time into utterance, they cannot be pre-computed; thus, when normalization is required, at run-time they must be calculated for each word in the vocabulary. However the normalization step is probably only needed for the sake of fair perplexity calculations, and not for speech recognition.

As the baseline we used a simple order 3 (trigram) trigram model, as implemented in SRILM (Stolcke, 2002), with the default parameter settings. The time-based adjustments were implemented as a wrapper around the function NgramLM::wordprobBO in the SRILM toolkit (Stolcke, 2002).

3.4. Training, Tuning, and Test Data

Following standard practice, we evaluate our various augmented language models using perplexity, a measure of the accuracy of predictions. As usual, we have the model assign probabilities to all words, but its success is judged only by the ability to assign high probabilities to the words that actually turn out to occur.

The training, tuning, and test data were all subsets of Switchboard. The training data was 1000 tracks, consisting of about 652K words. A separate set of 34K words was used as tuning data to determine the best value for the meta-parameters. The most important meta-parameter was k, specifying the importance given to the new information relative to the n-grams. All tokens were converted to lower case.

The test set consisted of 16 tracks from Switchboard, containing 10441 words and representing about 75 minutes of speech. For the experiments we limited the vocabulary to 5000 words, with other words treated as unknown; thus we made no attempt to predict them, and they were excluded from the perplexity computations, following a standard choice in language model evaluation. For evaluation purposes we also ignored sentence-end tags; this is because the utility of temporal and prosodic information for endpoint prediction is already well-known (Ferrer et al., 2003; Raux and Eskenazi, 2009), and we wanted to isolate the benefit for word prediction.

4. Features

Having shown a way to incorporate the information provided by a new feature in a language model in order to evaluate its utility, this section considers 23 features. Each feature was evaluated using the same models, and the same training, tuning, and test sets.

In general features chosen were among those that previous research indicates might be revealing of cognitive state and/or communicative intention. The rest of this section discusses each in turn, explaining the motivations, the implementation, and the benefits seen. First we discuss time into utterance and a related feature, then some prosodic features (which turned out to be by far the most informative), then features related to disfluencies, laughter, back-channeling, and time until end.

	best k value(s)	perplexity	benefit
baseline, $P_{trigram}$	–	107.766	–
with time into utterance, P_n	0.30	107.412	0.354
with time since other's end	0.30	107.425	0.341
with both, k retuned	0.35, 0.35	107.105	0.661

Table 3: Conditioning with Respect to Two Reference Events, and their Combination

4.1. Time into Utterance

Time into utterance, discussed above, did indeed bring useful information, as seen by the small perplexity improvement, as seen on the second line of Table 3. k was 0.3, chosen to give best performance on the tuning set.

These results were obtained with a model improved from that described earlier in two ways. First, the length of pause used for segmentation into utterances was 1.2 seconds, determined by optimization on the tuning set. (We also tried using the segmentation into utterances provided by the hand-labeled transcripts; this actually reduced the benefit seen, mostly because it raised the performance of the trigram baseline.) Second, bucket-based scaling was not applied if a word occurs at the start of an utterance, because in this position the probability is accurately modeled by the bigram <*s*> *word*: the fact that the word is also in bucket 0 brings no new information. As time-based scaling thus has nothing to offer such words, they are also excluded from training; specifically they are not included in the bucket 0 counts nor in the unigram counts, and thus they do not contribute to the computation of P_{tb} or anything else. We also experimented with varying the number of buckets and their widths, but this had only minor effects (Ward and Vega, 2010).

4.2. Time Since Other's End

While time into utterance is easy to compute, it is in a sense only a proxy for a deeper feature, time since the cognitive event of the initiation of the formulation process, which would probably be more predictive. This suggests several ways to improve this feature. For one, it would be possible to refine the notion of utterance-start, to require not just a preceding pause but also additional conditions such as presence of an uncontested turn exchange, however that might be defined. For another, it would be possible to use human-labeled utterance-start points, or start points infered by an algorithm trained on such data. However we opted for the simpler approach of conditioning on additional proxies.

The first such proxy is the time when the interlocutor ends his turn. Often this will be just before the time which the speaker starts vocalizing, but it may also be earlier or later, for example when a short overlapping turn by the interlocutor is followed by the speaker producing a *yeah* in mid-utterance. The specific variable used was the time since the most recent point at which the interlocutor ended a word and began a longish pause. Using pauses of at least 1.2 seconds as delimiters again gave the best performance.

Examining the ratios showed that the information given by conditioning on time since other's end did not merely duplicate that given by time into utterance. For example, that the probability of the word *so* is almost unrelated to time into utterance, but is higher in the first half second after the interlocutor has ended an utterance; and *you* and *yeah* are common after self-start, but almost unrelated to other-end.

We therefore wanted to combine the two sources of information. Subsequent sections will discuss this further, but here we just explain how this can be done and report the results for this feature pair. In fact it is easy: we tweak the trigram probabilities using both scaling factors, by adding an additional multiplicative factor to Equation 3. Using times since multiple reference events gives a sort of multi-layered representation of the speaker state at any time. Doing this gave an almost additive improvement, confirming that the two sources of information were more complementary than redundant. By increasing the values of k to 0.35 for each model performance was even better, as seen in Table 3. It seems that in some cases the two sources of information are not only largely non-redundant; they actually compensate for each others' weaknesses, enabling us to increase the weight of both.

4.3. Prosody in Language Modeling

The next four features all relate to the local prosodic context, so it is worth briefly discussing other approaches to the use of prosody in language modeling.

To date, most work in prosody for language modeling has focused on genres (notably radio broadcasts) and languages where lexical stress patterns are clearly realized or where prosody reveals syntactic structures (Shriberg and Stolcke, 2004a; Chen et al., 2007; Huang and Renals, 2007; Ananthakrishnan and Narayanan, 2007; Huang and Renals, 2007; Vicsi and Szaszák, 2010). In dialog, however, we expect that the effects of lexical and syntactic factors on the prosody to generally be obscured by the stronger effects of cognitive and interpersonal factors, such as delays while thinking of the right word and the management of who speaks when.

The exception, mentioned before (Stolcke et al., 1999), did use prosodic features related to cognitive factors, although, surprisingly, only durational factors were found to be of value.

Our choices of which specific prosodic features to use and how to compute them are in accordance with our goal of exploiting information related to cognitive states: the features are direct ones, in Shriberg's sense (Shriberg and Stolcke, 2004b), not hand-labeled nor inferred to match hand-labeled tags. Further, they are not syllable-aligned nor syllable-normalized; and they are computed over local contexts, not over entire utterances. In line with our aim of merely exploring the possibilities, many opportunities for tuning were passed up, but we did find the best window sizes for computing the features and then the best values for k.

In passing, it is worth observing that some work that adds prosodic information to language models does so not in order to help with the prediction of the upcoming words, as here, but just because it's sometimes easier to fit prosodic information into a language model than into an acoustic model. In the latter case, it is common to use prosodic features computed wholely (or partially (Stolcke et al., 1999)) over the word to "predict." There is nothing wrong with this, but this difference should be kept in mind when comparing performance benefits.

previous speaking rate	characteristic words . . .	. . . uncharacteristic words
fast	sixteen, carolina, o'clock, kidding, forth, weights, familiar, half, science, process, careful, matter, grand, doubt, talking, role . . .	. . . hm, uh-huh, ah, huh
middling	direct, wound, mistake, mcdonald's, likely, wears, troops, term, repairs, purchased, lawyer, immigration, guard, director, minimum . . .	. . . uh-huh, hi, um-hum
slow	goodness, gosh, agree, bet, let's, uh, god, um, grew, huh-uh, although, neat, either, definitely, true, am, bye-bye, unless, thank . . .	. . . experience, yourself, ago
(none)	um-hum, uh-huh, hum, hm, oh, yep, yeah, wow, huh, yes, ah, right, okay, well, exactly, no, sure, which . . .	. . . guess, know, mean, lot

Table 4: Characteristic and Uncharacteristic Words in Different Speaking-Rate Contexts

4.4. Speaking Rate

We first considered speaking rate, as likely to indicate degree of preparation and confidence, and because word durations are strongly affected by frequency and predictability (Bell et al., 2009; Bradlow and Baker, 2010). Each token in the corpus was characterized in terms of speaking rate: tokens less than 0.89 of the average duration for that word were considered fast, more than 1.11 of the average duration slow, and the rest middling. Each token was then put in one of four buckets, after-slow, after-middling, after-fast, or after-silence, depending on the duration of the previous word, if any. These characterizations were done from the transcriptions, without reference to the actual speech signal.

We then calculated which words tended to occur in which contexts: Table 4 shows the most characteristic and uncharacteristic. Examining the words in each category suggests some patterns. Common after fast regions (words of relatively short duration) are high-content words, especially place names and numbers. Common after slow regions (words of relatively long duration) are assessments, disfluency markers, social expressions (*bye-bye, thank [you]*) expressions of belief (*definitely, unless, well, yes, [of] course, but, consider, absolutely, okay, must, generally, certainly, totally*), and the word *I*.

On the tuning data predictions were improved for words in the fast and slow contexts, but not in the middling rate context, so we dropped words in middling-rate contexts from the model. This gave the best single-feature perplexity improvement, 2.771 points, at a k value of 0.99. (This and subsequent results as summarized in Table 5 below.) Overall, the words that gave the maximum benefit were *um-hum, yeah, uh, oh, I, uh-huh*, and *you*, all of which are more common in slow contexts. Among the various possible normalization schemes (Batliner et al., 2001), we also tried normalizing with respect to the speaker's overall rate, such that a word would be characterized as fast or slow with reference to the average for that particular speaker, however this was not advantageous.

Because these results were obtained from human-labeled word durations, and the duration estimates available to a speech recognizer will not be so accurate, we also tried conditioning on a purely acoustic proxy for speaking rate. Specifically, using the sum of the absolute values of the differences in energy between adjacent frames, normalized by the difference between the average speaking volume and the average silence volume, gave a very rough approximation to syllable rate. We computed this, over regions of size 325ms immediately previous to the onset of the word to predict. Using this we again classified each token as after-none (after a period with very little variation in energy), after-slow, after-middling, or after-fast (after a period with a lot of variation in energy). The results were not as good as those obtained using the hand-labeled durations, suggesting that the accuracy of the rate estimates is important. Use of a better rate estimator, perhaps mrate (Morgan and Fosler-Lussier, 1998), seems needed.

4.5. Volume

We considered volume, as likely to indicate states such as engagement or dominance. Volume was speaker-normalized. Specifically, for each dialog side we look at a large sample of regions, and used an Expectation Maximation algorithm to find the mean volume of silence regions and the mean volume of speech regions. Regions with an energy closer to the silence mean than to the speaking mean were considered "silent," those with an energy within one standard deviation of the mean as "moderate," those less as "quiet," and those more as "loud." Each word was then associated with the loudness label over the region immediately preceding it.

Best performance on the tuning data was obtained when volume was computed over windows 50ms wide, shorter than we had expected. Common after quiet regions are expressions of belief (*[I] bet, [I] know, y[ou know], true*), of types and degrees of belief (*although, mostly, definitely, might, usually, tend, looks, guess*), and clause connectives (*well, then*). Common after moderate-volume lead-ins are the tail ends of multi-word expressions (*[and so] forth, [San] Francisco, [New] Hampshire, [to some] extent*). Common after loud lead-ins are general content words, and, to a lesser extent, words pronounced while laughing. Other examples are given in (Ward and Vega, 2009). As always, the interpretation of such tendencies is not clear-cut. Some, but not all, appear to relate to patterns of lexical stress. Others seem to reflect cognitive states and/or communicative situations: for example, the tendency for expressions of belief to come after low-volume regions may reflect a cognitive state or a communicative strategy of preceding important words with a quiet lead-in, to give them more impact. Regardless of what the correct interpretation might be, a fair perplexity improvement was obtained.

4.6. Pitch Height and Pitch Range

We considered pitch height, as a possible indication of involvement and local dominance. Specifically we used the median pitch over the 150ms immediately preceding the onset of the word to predict, and then characterized this as low, medium, high, or no-pitch, depending on the position of this median relative to the 30th percentile and 70th percentile pitch levels computed over the entire track.

A fair perplexity improvement was obtained. Common words after regions of low pitch height were low-frequency content words, such as *privacy, body, forth, teaching, retirement, oil* and number words; after regions of medium height more common nouns; and after high pitch regions laughter and words pronounced while laughing, and emotion words such as *admit, surprised, kidding*, and *incredible.*

Pitch range was included as a features possibly indicative of interest and emotion. Reliability being an issue, we discarded the top and bottom pitchpoints in each 225ms region and used the ratio between the second-highest and second-lowest pitch points as our measure of range. This value was then compared the range to the maximum range value seen in a large sample of regions across the entire track. Regions with a range less than 0.3 of this maximum were considered to have a narrow range; those over 0.5 to have a wide range.

A fair perplexity improvement was obtained. Common words after narrow-range regions were generally low-frequency words; after moderate range regions many one-syllable words, and after high range regions positive emotion words such as *wonderful, fun, family, great, best, nice, family* and *bought.*

4.7. Time Since Fillers, Word Fragments and Laughter

In Switchboard many utterances start with filler words, and sometimes it seems that the "real" start of the utterance comes at the point when the fillers end and the content words begin. Without specifically identifying filler sequences, we just conditioned on the time since the most recent filler found in the transcript, which for convenience we approximated as all occurrences of the words *uh*, *yeah*, *um*, *well*, *right*, *oh*, *[vocalized-noise]*, *okay*, *uh-huh*, *huh*, and *um-hum.*

Given the importance of disfluency modeling (Stolcke et al., 1999), we also examined word fragments. Word fragments are common in Switchboard, and may be a good proxy for disfluency events, possibly indicating in particular when a speaker doesn't have his utterance planned out far ahead, with implications for what words may appear over the next few seconds. We conditioned on time since the onset of the most recent fragment, identified as words that were transcribed as only partially pronounced, for example *ap[-ple]* for *apple.* We also conditioned on time since last filler/fragment by the interlocutor. In every case the benefits were small.

We also considered laughter, as laughter is often a salient dialog event, although generally one with multiple interpretations. Since laughter is a relatively rare occurrence, when using this reference event most of the corpus fell into the later buckets, which may be a reason why the benefits were small. The comparative rarity of laughter events may also be why we saw a detriment in one case: some idiosyncratic similarities between the laughter in the tuning data and in the training data probably resulted in best performance at a high (0.8) weight for k, but that weight was apparently too strong and hurt performance in the test data.

4.8. Time Since Back-Channels and Low-Pitch Regions

We also considered back-channels. In general a person who who has just produced a back-channel is indicating the intention to continue in a listener role, and is probably unlikely to say anything contentful soon. On the other side, a person who has just received a back-channel from the other person is probably likely to continue speaking and perhaps become more fluent and contentful. Thus we conditioned on the onset of back-channels, specifically tokens labeled *uh-huh* or *um-hum* in the transcription (Hamaker et al., 1998), as a reference event. Fair results were seen.

Because speakers of English frequently invite back-channel feedback from the interlocutor by producing low pitch regions (Ward and Tsukahara, 2000), we thought that the recent presence of such a region might also provide clues as to what sort of thing they might be preparing to say next. Fairly good results were obtained.

4.9. Time-Until Features

Having obtained good results with time into utterance, we also tried time until end of utterance, reasoning that a speaker planning to finish in a second or two is likely to be in a different cognitive state from one intending to keep talking for a while. This feature, unlike most, was computed as the time from the *end* of the word of interest. Common just before the end were words such as *um-hum, vocalized noise*, and *uh-huh*; the least common were *where, an* and *years*.

We also tried time until interlocutor's next start of utterance. Just before this, the most common were *laughter, but*, and *so*, presumably as marking easy places for the interlocutor to jump in and take the turn; the least common were *um-hum, uh-huh*, and *oh*. For time until own next low pitch region, the most common words were *um-hum, uh-huh*, and *oh*, and the least common were *think, have*, and *don't*; presumably because a speaker in a fluent patch and about to introduce content is unlikely to solicit a back-channel soon.

For all these features the predictive value was high. However it is worth noting that, since these features involve future information, they would not be useful for on-line speech recognizers.

4.10. Discussion

Table 5 shows the results for all the features discussed. From this we can make a few observations. First, it seems that the most informative features were the prosodic ones. Second, the optimal windows for the prosodic features were surprisingly short, probably in part because the cognitive states of interest are often quite short and their effects immediate, and probably in part because they are capturing more lexical stress information than we had expected. Third, the time-since features refering to the behavior of the interlocutor are often as informative as those relating to the behavior of the speaker himself.

5. Towards the Application of Such Features

Although far from satisfied with the results so far, as discussed later, we went on to explore how the information in the features could be combined and used in a speech recognizer.

5.1. Combined Models

Wanting to see how well the features would perform together, we took the 8 best-performing features and combined them. Although the features are clearly not independent, we chose as first step to treat them as such, combining the tweaks by simply multiplying

Feature	Best k in Isolation	Benefit in Isolation
time into utterance	0.30	0.354
time since interlocutor's end	0.40	0.364
speaking rate (from previous word duration)	0.99	2.771
” (over the previous 325ms.)	0.55	1.136
volume (over the previous 50ms.)	0.49	2.651
pitch height (over the previous 150ms.)	0.60	2.046
pitch range (over the previous 225ms.)	0.55	1.741
onset of own last filler	0.25	0.104
end of own last filler	0.35	0.076
onset of other's last filler	0.25	0.151
end of other's last filler	0.35	0.101
time since own last fragment	0.35	0.130
time since other's last fragment	0.40	0.020
time since own laughter onset	0.50	0.082
time since own laughter end	0.80	–0.038
time since other's laughter onset	0.70	0.089
time since other's laughter end	0.55	0.022
time since own back-channel	0.35	0.132
time since other's back-channel	0.40	0.149
time since own low pitch region	0.50	0.226
time since other's low pitch region	0.45	0.218
time before end of utterance	0.30*	0.705
time before other's start	0.30*	0.176
time before low-pitch region	0.30*	0.780

Table 5: Perplexity Improvements on Switchboard. The “benefit” is the perplexity decrease relative to the baseline model; the third column shows this for each feature with the k value shown in the second column. * unoptimized

Feature	k in the Combined Model
speaking rate (over the previous 325ms.)	0.45
volume (over the previous 50ms.)	0.45
pitch height (over the previous 150ms.)	0.30
pitch range (over the previous 225ms.)	0.10
time into utterance	0.40
time since interlocutor's end	0.25
time since own low pitch region	0.25
time since other's low pitch region	0.25

Table 6: Features and Weights used in the Combined Model for Switchboard.

Model	Benefit
fully automatic	4.788
" some redundancy removed	5.001
rate based on labels	5.741
" some redundancy removed	5.898

Table 7: Perplexity Improvements on Switchboard for the combined models.

all of their contributions (Equation 4), that is, by using them all as simultaneous tweaks on the trigram probabilities. We then found good k values for this combined model, as seen in Table 6, by hill-climbing. The slow pace of this process was the reason for only using 8 features. The resulting model has nominally 560,000 parameters (24 time slices x 4 reference events plus 4 categories x 4 prosodic features, for each of 5000 words), of which about 42,000 are not equal to 1.0.

Table 7 shows the results. The first line shows the results when rate was estimated using our rough estimate of speaking rate. Over the whole test set, 63% of the occurring words had their probability estimates improved by this model. Examination of why some words received penalties rather than benefits revealed several patterns (Ward et al., 2010b), most saliently the occasional application of multiple redundant (non-independent) penalties from the prosodic features for words after pauses. Reducing this redundancy by turning off prosodic weighting for words at the start of utterances gave a small improvement, as seen in line two of the Table.

Although, as noted above, computing the perplexity benefits using rate as estimated from hand-labeled durations is not reasonable for estimating potential value for speech recognition, curiosity led us to measure the benefit of using this anyway. Using this in a combined model, gave a better perplexity reduction. Again a version where no prosodic tweaks were applied to utterance-initial words did better, giving, as seen in the fourth line, a perplexity of 101.868, a 5.5% decrease.

We also measured the perplexity benefit on a second corpus of spontaneous dialogs, the

Feature	Best k in Isolation	Benefit	k in the Combined Model
speaking rate (over the previous 50 ms.)	0.35	0.20	.30
volume (over the previous 200 ms.)	0.55	1.91	.35
pitch height (over the previous 125 ms.)	0.60	2.52	.45
pitch range (over the previous 50 ms.)	0.45	0.84	.15
combined		3.271	

Table 8: Perplexity Improvements on Verbmobil.

German Verbmobil-II corpus (Jekat et al., 1997). This corpus contains 14K utterances with a total of 167K words (5K word types). We split the corpus into training, development/tuning and test sets, taking care that audio from one dialogue would not end up in more than one of the sets. This resulted in a test set of 1101 utterances (14K words). Baseline perplexity, determined as for Switchboard, was 106.271.

For simplicity, we built a model using only the four prosodic features. As before, these were normalized per dialog side, thus utterances from one speaker in a dialog were examined in order to determine that speaker's typical volume, pitch height, etc. Wishing to reduce the redundancy among the features, we chose the window sizes iteratively, using a greedy algorithm: thus we first found the window size for volume that gave the greatest perplexity reduction on the tuning set, then the window size for pitch height that gave the greatest perplexity reduction when used together with volume, and so on for pitch range and speaking rate. This was done with all k values fixed at 0.3; then we did hillclimbing to find a better set of k. As before, no tweaks were applied to utterance-initial words. The resulting window sizes and k values are seen in Table 8. This combined model gave a perplexity of 103.000, a 3.1% reduction over the baseline.

5.2. Use in a Speech Recognizer

Although the perplexity improvements obtained were not large enough to suggest a significant benefit for speech recognition, we went ahead and tried anyway. Not having good acoustic models for English at hand, we tried it on the Verbmobil corpus, using the Sphinx-4 (Walker et al., 2004) recognizer with simple one-pass decoding without subsequent hypothesis re-ranking. The acoustic model and (trigram) language model were trained on the training and development sets, giving a baseline word error rate of 34.6%.

To incorporate non-lexical context information, we extended Sphinx's trigram language model to lookup and apply the prosodic tweak to the trigram score whenever the language model was called for a word hypothesis. No normalization was done, that is, we used P_{bs} instead of P_n, to avoid unnecessary computation. As an implementation detail, the prosodic features were pre-computed for each audio file, and then read in by Sphinx. This was done because we had already set up a pipeline for prosody processing which worked outside of Sphinx and wanted to keep it for simplicity. The same results could be achieved if the prosodic processing was done online during recognition. Although it would be possible to use the recognizer's result so far to infer the speaking rate, we kept with our simple acoustic measure.

Despite the perplexity reduction seen (3.1%) the word error increased, from 34.6% to 34.7%. Among the many possible reasons for this (Chen et al., 1998), one likely factor is that the model is not trained to give probability estimates at times which are not word boundaries, although in a recognizer it has to provide such estimates. This problem is exacerbated by the fact that the alignments of the word hypotheses (even the ultimately correct ones) in the recognizer are not the same as those seen in the hand-labeled data used in training. In fact, we observed an average offset of 11 ms, with a standard deviation of 61 ms, for these hypotheses, which could be significant, given the small size of the windows used for some of the prosodic features.

6. Summary and Directions for Future Work

Our exploration, by design unguided by knowledge of previously examined theoretical constructs and by considerations of ease of modeling, indeed led to the discovery of new dependencies, some of which appear not to overlap those handled by existing models, thus enriching our inventory of potentially useful predictive features. We have also shown that this information can be used in language models, giving perplexity reductions of 5.5% on Switchboard and 3.1% on Verbmobil II, both relative to trigram baselines. These findings suggest that our four principles of Section 2 do indeed describe a promising approach to better language modeling for dialog. However at this point the quantitative benefits are modest at best, and no speech recognition improvement was seen.

Among the many possible extensions and improvements, a few seem especially promising. One would be to use more and better features, for example, prosodic features computed over additional window sizes. Another would be to treat the features as continuous, rather than discretizing them, which could reduce the number of parameters in the model and reduce sparseness problems. Another would be to avoid treating the features as independent, instead perhaps using a few principal components as a concise characterization of the relevant state of a dialog participant at any moment. Finally, the modeling method (Section 3) should probably be replaced by one that can be combined with n-gram information in a more principled way. This could be approached by reformulating using probabilities and information theory, or perhaps integrating non-lexical predictors into a more general framework (Bengio et al., 2003; Xu and Jelinek, 2007).

Although clearly not ready yet for speech recognition, language models incorporating non-lexical features may be useful for other purposes, for example language generation and speech synthesis. Such models may also be of use for predictions that go beyond words, for example to predict the exact timings of the words and non-lexical vocalizations, the nuances of their pronunciations, their pitch and energy contours, and ultimately also gestures. Doing so could be a way to systematize and improve turn-taking (Ward et al., 2010a) and various forms of adaptation in dialog.

Beyond practical applications, the dynamics of participation in spoken dialog are a topic of great scientific interest: many researchers have pointed out that dialog behaviors offer a unique window into human cognition and human social interactions (Yngve, 1970; Sacks et al., 1974; Clark, 2002; Pickering and Garrod, 2004; Levinson, 2006; O'Connell and Kowal, 2008). Despite the intrinsic interest of these topics, and the notion that "humans are 'designed' for dialogue rather than monologue" (Garrod and Pickering, 2004), to date most modeling work, in psycholinguistics and other fields, has focused either on production alone or comprehension alone. Although language modeling has in the past been seen as a purely technical, applied enterprise, its techniques for making and evaluating predictions may provide a valuable addition to the set of tools for understanding human behavior. Our finding here, that word probabilities vary with shallow cognitive states, may open a path towards to the development of deeper models of the cognitive processes underlying dialog.

References

Ananthakrishnan, S., Narayanan, S., 2007. Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework. In: ICASSP. pp. 873–876.

Bard, E. G., Aylett, M. P., Lickley, R. J., 2002. Towards a psycholinguistics of dialogue: Defining reaction time and error rate in a dialogue corpus. In: Bos, J., Foster, M., Matheson, J. (Eds.), EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue. pp. 29–36.

Barsalou, L. W., Breazeal, C., Smith, L. B., 2007. Cognition as coordinated non-cognition. Cognitive Processing 8, 79–91.

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke, V., Niemann, H., 2001. Duration features in prosodic classification: Why normalization comes second. In: ISCA Workshop on Prosody in Speech Recognition and Understanding.

Beebe, B., Badalamenti, A., Jaffe, J., Feldstein, S., et al., 2008. Distressed mothers and their infants use a less efficient timing mechanism in creating expectancies of each other's looking patterns. Journal of Psycholinguistic Research 37, 293–307.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., Jurafsky, D., 2009. Predictability effects on durations of content and function words in conversational English. Journal of Memory and Language 60, 92–111.

Bellegarda, J. R., 2004. Statistical language model adaptation: review and perspectives. Speech Communication 42, 93–108.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. Journal of Machine Learning Research 3, 1137–1155.

Bradlow, A. R., Baker, R. E., 2010. Variability in word duration as a function of probability, speech style and prosody. Language and Speech(in press).

Brennan, S. E., Hulteen, E. A., 1995. Interaction and feedback in a spoken language system: A theoretical framework. Knowledge-Based Systems 8 (2–3), 143–151.

Campbell, N., 2007. On the use of nonverbal speech sounds in human communication. In: Esposito, A., et al. (Eds.), Verbal and Nonverbal Communicative Behaviours, LNSI 4775. Springer, pp. 117–128.

Chen, K., Hasegawa-Johnson, M., Cole, J., 2007. A factored language model for prosody-dependent speech recognition. In: Grimm, M., Kroschel, K. (Eds.), Robust Speech Recognition and Understanding. I-Tech, pp. 319–332.

Chen, S. F., Beeferman, D., Rosenfeld, R., 1998. Evaluation metrics for language models. In: DARPA Broadcast News Transcription and Understanding Workshop.

Clark, H. H., 1996. Using Language. Cambridge University Press.

Clark, H. H., 2002. Speaking in time. Speech Communication 36, 5–13.

Ferrer, L., Shriberg, E., Stolcke, A., 2003. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In: ICAASP.

Fujisaki, H., 2008. In search of models in speech communication research. In: Interspeech. pp. 1–10.

Garrod, S., Pickering, M. J., 2004. Why is conversation so easy? Trends in Cognitive Sciences 8, 8–11.

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

Goffman, E., 1981. Response cries. In: Goffman, E. (Ed.), Forms of Talk. Blackwell, pp. 78–122, originally in *Language* 54 (1978), pp. 787–815.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., Morency, L.-P., 2006. Virtual rapport. In: 6th International Conference on Intelligent Virtual Agents. pp. 14–27.

Hamaker, J., Zeng, Y., Picone, J., 1998. Rules and guidelines for transcription and segmentation of the Switchboard large vocabulary conversational speech recognition corpus, version 7.1. Tech. rep., Institute for Signal and Information Processing, Mississippi State University.

Huang, S., Renals, S., 2007. Modeling prosodic features in language models for meetings. In: Popescu-Belis, A., Renals, S., Bourlard, H. (Eds.), Machine Learning for Multimodal Interaction IV (LNCS 4892). Springer, pp. 191–202.

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

Jaffe, J., 1978. Parliamentary procedure and the brain. In: Siegman, A. W., Feldstein, S. (Eds.), Nonverbal Behavior and Communication. Lawrence Erlbaum Associates, pp. 55–66.

Jahr, E., Eldevik, S., 2007. Response variability and turn taking in cooperative play. Journal of Speech and Language Pathology 2, 190–194.

Jekat, S., Scheer, C., Schultz, T., 1997. VMII Szenario I: Instruktionen für alle Sprachstellungen. Tech. Rep. VM-Techdoc 62, Universität Hamburg, LMU München, Universität Karlsruhe.

Jelinek, F., 1997. Statistical Methods for Speech Recognition. MIT Press.

Ji, G., Bilmes, J., 2004. Multi-speaker language modeling. In: HLT.

Ji, G., Bilmes, J., 2010. Jointly recognizing multi-speaker conversations. In: ICASSP 2010.

Levinson, S. C., 2006. On the human 'interaction engine'. In: Enfield, N. J., Levinson, S. C. (Eds.), Roots of Human Sociality. Berg, pp. 39–69.

Ma, K. W., Zavaliagkos, G., Meteer, M., 2000. Bi-modal sentence structure for language modeling. Speech Communication 31, 51–67.

Macrae, C. N., Duffy, O. K., Miles, L. K., Lawrence, J., 2008. A case of hand waving: Action synchrony and person perception. Cognition 109, 152–156.

Morgan, N., Fosler-Lussier, E., 1998. Combining multiple estimators of speaking rate. In: ICASSP. IEEE, pp. 721–724.

O'Connell, D. C., Kowal, S., 2008. Communicating with One Another: Toward a Psychology of Spontaneous Spoken Discourse. Springer.

Petukhova, V., Bunt, H., 2009. The independence of dimensions in multidimensional dialogue act annotation. In: NAACL-HLT.

Pickering, M. J., Garrod, S., 2004. Toward a mechanistic psychology of dialog. Behavioural and Brain Sciences 27, 169–190.

Raux, A., Eskenazi, M., 2009. A finite-state turn-taking model for spoken dialog systems. In: NAACL HLT.

Sacks, H., Schegloff, E. A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.

Sebanz, N., Bekkering, H., Knoblich, G., 2006. Joint action: Bodies and minds moving together. Trends in Cognitive Sciences 10, 70–76.

Shriberg, E., Stolcke, A., 2004a. 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. Springer-Verlag, pp. 105–114.

Shriberg, E. E., Stolcke, A., 2004b. Direct modeling of prosody: An overview of applications in automatic speech processing. In: Proceedings of the International Conference on Speech Prosody. pp. 575–582.

Stolcke, A., 2002. SRILM – an extensible language modeling toolkit. In: Proc. Intl. Conf. on Spoken Language Processing.

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

Streek, J., Jordan, J. S., 2009. Projection and anticipation: The forward-looking nature of embodied communication. Discourse Processes 46, 93–102.

Vicsi, K., Szaszák, G., 2010. Using prosody to improve automatic speech recognition. Speech Communication 52, 413–426.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J., 2004. Sphinx-4: A Flexible Open Source Framework for Speech Recognition. Tech. Rep. SMLI TR2004-0811, Sun Microsystems Inc.

Ward, N., Tsukahara, W., 2000. Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32, 1177–1207.

Ward, N. G., Fuentes, O., Vega, A., 2010a. Dialog prediction for a general model of turn-taking. In: Interspeech.

Ward, N. G., Vega, A., 2009. Towards the use of inferred cognitive states in language modeling. In: 11th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 323–326.

Ward, N. G., Vega, A., 2010. Studies in the use of time into utterance as a predictive feature for language modeling. Tech. Rep. UTEP-CS-2x.

Ward, N. G., Vega, A., Novick, D. G., 2010b. Lexico-prosodic anomalies in dialog. In: Speech Prosody.

Ward, N. G., Walker, B. H., 2009. Estimating the potential of signal and interlocutor-track information for language modeling. In: Interspeech. pp. 160–163.

Xu, P., Jelinek, F., 2007. Random forests and the data sparseness problem in language modeling. Computer Speech and Language 21, 105–152.

Yngve, V., 1970. On getting a word in edgewise. In: Papers from the Sixth Regional Meeting of the Chicago Linguistic Society. pp. 567–577.