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 features. Using perplexity as the metric, the most informative of these included recent speaking rate, volume, and pitch, and time until end of utterance. Using simple combinations of such features to augment trigram language models gave up to a 8.4 % perplexity benefit on the Switchboard corpus, and up to a 1.0 % relative reduction in word error rate (0.3 % absolute) on the Verbmobil II corpus.

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

[☆] We thank Shreyas Karkhedkar, Nisha Kiran, Shubhra Datta, Gary Beverungen, and Justin McManus for contributing to the code and analysis; David Novick, Olac Fuentes and many thoughtful reviewers for Speech Communication, Interspeech, ASRU, and the National Science Foundation 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, by the US Army Research, Development and Engineering Command, via a subcontract to the USC Institute for Creative Technologies, and by a DFG grant in the Emmy Noether program.

Email addresses: `nigelward@acm.org` (Nigel G. Ward), `avega5@miners.utep.edu` (Alejandro Vega), `mail@timobaumann.de` (Timo Baumann)

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

have been seen to correlate with more empathy and success in interactions (Gratch et al., 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 language 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 by using this as inspiration to try models which reflect the fact that spoken language is created by human minds and for human minds. This paper accordingly 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. Sections 4 and 5 examine several dozen non-lexical features, related to timing and prosody, and shows that many do provide useful information, as measured by perplexity reductions. Section 6 shows how a simple combination of such features can be used to benefit speech recognition. Finally, Section 7 summarizes and notes directions for future work.

2. Time, States, and Events in Dialog

One specific inspiration for our modeling strategy is 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 informative.

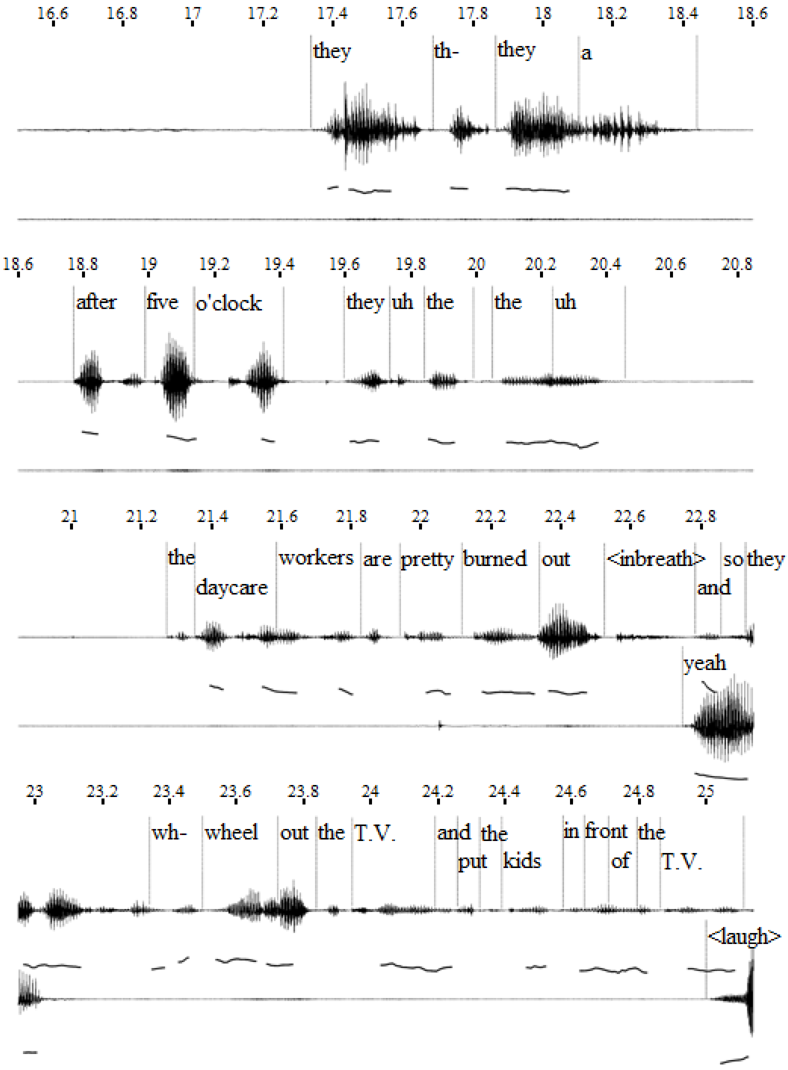

Figure 1: Conversation Fragment. Each of the four strips includes a timeline and two rows, one per speaker. Each row includes a transcription, the signal and the pitch. This fragment occurred after some talk about television-watching habits and effects on children. Audio for this clip is available at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

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 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, all of which used the Switchboard corpus (Godfrey et al., 1992).

Stolcke, Shriberg and colleagues (Stolcke et al., 1999) modeled 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 if they were additional words in the word sequence. Since

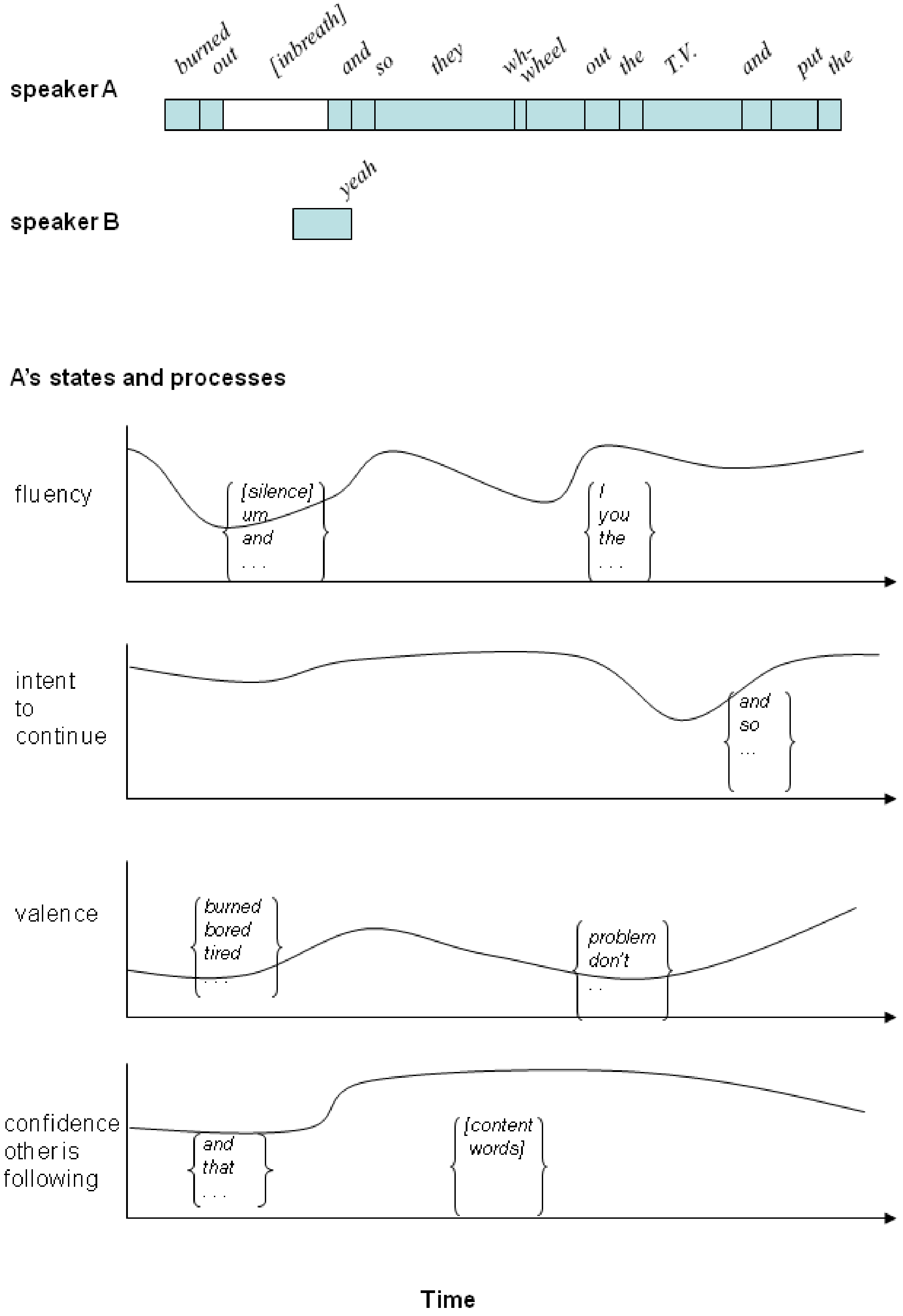

Figure 2: A fanciful image of the aspects of the cognitive state of the speaker in Figure 1 from 22 to 24.5 seconds. Each curve indicates the varying intensity over time of a cognitive state, process, or need. The words in curly braces suggest some of words more likely to be produced while in that state.

this formulation fits easily into the n-gram framework, future word probabilities can be conditioned on these events. The hidden events were detected using both lexical and prosodic context. Incorporated into a speech recognizer the model gave a 0.9 % absolute (2 % relative) decrease in word error rate.

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 for multiple cognitive processes operating in parallel, including not only those of formulating and speaking but also those of hearing and comprehending (Yngve, 1970; Jaffe, 1978; Bard et al., 2002).

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 perplexity reductions from 19 % to 36 % and 0.2–0.3 % decreases (0.8 % relative) 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 (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." More features will then be described and evaluated in Sections 4 and 5.

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 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, defined as sequences of words delimited by at least 1.2 seconds 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:

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

To explore 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

In an attempt to connect the observations back to our ideas about the underlying cognitive processes, we examined the words which were exceptionally common and exceptionally uncommon in various time buckets. We observed several patterns which do seem to reflect such processes. For example, low frequency words, typically content-rich words, are relatively more common later in utterances, and this 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 word *know* grows in frequency over time, being less than 1.4 % over the first 5 seconds but over 1.8 % after 10 seconds, perhaps because of the time it takes to reason about knowledge states. Interestingly 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

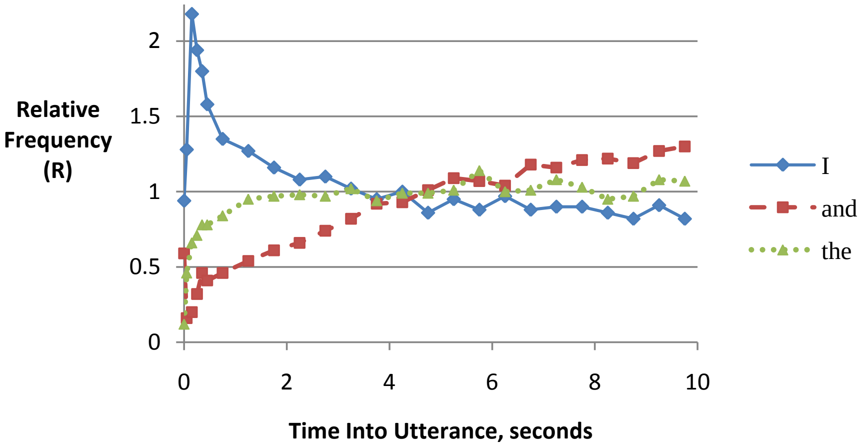

Figure 3: Ratio of Frequency in Specific Time Buckets to Overall Frequency (R) versus time into utterance, for the three most common words in Switchboard. The rightmost points represent the range 9.5 seconds and up.

also interesting: it occurs twice as often near the start of utterances as elsewhere, peaking at around 0.2 seconds in, and it is far more common initially than *you*, although 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 (Ward and Vega, 2008).

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 interested less in its value in isolation than in whether it proves information which is complementary to that provided by existing methods. While the information given by time into utterance could be used to augment any language model, we illustrate its use in combination with a trigram model.

Our first combination attempt used 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 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 the number 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, 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.

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

We choose not to apply bucket-based scaling 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. Zero-offset words are similarly excluded for all the other features discussed later.

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.

To deal with the inevitable sparseness issue we only individually modeled the most frequent 1000 words. Words outside this shortlist were modeled as a class. Doing this is reasonable because, as such words are generally content words, they can be expected to have similar patterns of occurrence.

Finally there is a normalization step to ensure that all the probabilities across the vocabulary add to 1 in each case. 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:

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

The normalization step is necessary in order to enable the perplexity calculations, which require true probabilities, not likelihoods. Since the combined probability estimates depend on the preceding words as well as the time into utterance, they cannot realistically be pre-computed, as they will differ for every combination of lexical context and bucket. Normalization therefore requires all the values to be calculated for each word in the vocabulary, at run-time. However the normalization step is probably only needed for the perplexity calculations, not for speech recognition, as discussed below. Finally, it is necessary to note that the normalization is done so as to ensure not only that the probabilities across the vocabulary add to 1 in each case, but also that the probability assigned to `</s>`, the end of sentence token, is unchanged from that assigned by the baseline model. That is, we ensure that our model does not devote more of the probability mass to words than does the baseline model, for the sake of fair comparisons.

The specific model we used as the baseline is a standard order 3 (trigram) model, namely the one implemented in the SRILM toolkit (Stolcke, 2002), run with the default parameter settings, which include Katz' back-off with Good-Turing discounting. The time-based adjustments are implemented as a wrapper around the function NgramLM::wordprobBO in SRILM.

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 981 tracks, consisting of about 80 hours of speech and 650K words. A separate set of

35K 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 45 tracks from Switchboard, containing 28K words and representing about 4 hours 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 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) so a substantial perplexity reduction could be obtained, but better prediction of sentence-ends has no value for speech recognizers operating on inputs that have already been segmented by a pre-processing step. The baseline perplexity on the test set was 109.449. Results below are reported in terms of percentage reduction in perplexity relative to this baseline.

4. Temporal Features

This section examines a number of potential reference events in dialog, examining their utility for language modeling using the methods developed above. These were chosen from among those that previous research indicates might be revealing of cognitive state and/or communicative intention. First we discuss utterance starts and ends then features related to disfluencies, laughter, and back-channeling.

4.1. Utterance Start and End

Time into utterance, our example from above, did indeed bring useful information, as seen by the 0.31 % perplexity improvement, also seen as the first number in Table 1. Again, utterance boundaries are determined by looking for 1.2 second periods of silence. This value was chosen as the one that gave best performance for the baseline model. It was also, conveniently, the best for conditioning on this feature. We also experimented with varying the number of buckets and their widths, but this had only minor effects (Ward and Vega, 2010).

Time into utterance is really 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. It might be possible to better approximate this by refining the notion of utterance-start, to require not just a preceding pause but also, for example, the presence of an uncontested turn exchange, however that might be defined. 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

Perplexity Improvements with Conditioning on Time Since

		Start of		End of	
Utterance	by self	0.31 %	.5	0.09 %	.3
	by interlocutor	0.36 %	.6	0.43 %	.5
Filler	by self	0.10 %	.3	0.23 %	.4
	by interlocutor	0.24 %	.6	0.17 %	.5
Fragment	by self	0.04 %	.4	0.04 %	.3
	by interlocutor	0.04 %	.6	0.02 %	.3
Back-Channel	by self	0.05 %	.3	0.07 %	.4
	by interlocutor	0.20 %	.7	0.18 %	.6
Pure Laughter	by self	0.07 %	.5	0.10 %	.5
	by interlocutor	0.12 %	.8	0.01 %	.8
Laughed Word	by self	0.06 %	.4	0.05 %	.5
	by interlocutor	0.04 %	.6	0.02 %	.4
Low-Pitch Region	by self	0.65 %	.6	0.12 %	.4
	by interlocutor	0.15 %	.4	0.19 %	.3

Table 1: Perplexity Improvements on Switchboard, conditioning on time *since* various events. The "benefit" is the perplexity decrease as a percentage relative to that of the baseline model. The second number in each column is the best weight (k).

Perplexity Improvements with Conditioning on Time Until

		Start of		End of	
Utterance	by self	0.40 %	.6	2.09 %	.7
	by interlocutor	1.00 %	.7	0.35 %	.6
Filler	by self	0.80 %	.6	—	–
	by interlocutor	0.57 %	.7	0.71 %	.7
Fragment	by self	0.00 %	.4	—	–
	by interlocutor	0.03 %	.4	0.04 %	.3
Back-Channel	by self	0.14 %	.6	—	–
	by interlocutor	0.24 %	.7	0.34 %	.7
Pure Laughter	by self	0.11 %	.8	—	–
	by interlocutor	0.04 %	.7	0.21 %	.7
Laughed Word	by self	0.04 %	.7	—	–
	by interlocutor	0.02 %	.5	0.01 %	.5
Low-Pitch Region	by self	0.92 %	.6	1.25 %	.7
	by interlocutor	0.15 %	.5	0.18 %	.5

Table 2: Perplexity Improvements on Switchboard, conditioning on time *until* various events.

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, 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, which we did by applying the scaling factors for both models together. 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. Specifically, with k values of .4 and .45, a perplexity reduction of 0.7 % was obtained, as seen below in Table 5.

For the sake of completeness, we completed the paradigm; thus the first four results in Table 1 show the benefits for conditioning on time since start/end of utterance by self/other.

We also tried time-until features, obtaining the results shown in Table 2. Noteworthy are the good results found with time until end of utterance by self. Common words just before the end included *um-hum, [vocalized-noise]*, and *uh-huh*; relatively uncommon ones included *where, an* and *years.* The power of this feature could be a reflection of the fact 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. Alternatively one could see it as reflecting the tendency of speakers to signal by their word choice when their utterance is about to end.

Another strong feature is time until the start of an utterance by the interlocutor. Just before this, the most common words included *[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.*

It is worth noting that, since the time-until features involve future information, they would probably not be useful for dialog systems, although they would for offline applications such as transcription, voice search, and wordspotting.

4.2. 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; thus the occurrence of a filler can be taken as another proxy for a cognitive starting point of formulation. Without specifically identifying filler sequences, we just conditioned on the time relative to the closest filler in the transcript, which for simplicity 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 value of disfluency modeling (Stolcke et al., 1999), we also examined word fragments. Word fragments are common in Switchboard, and may be a 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 closest fragment, identified as words that were transcribed as only partially pronounced, for example *ap[-ple]* for *apple*.

We also considered laughter, as laughter is often a salient dialog event, although generally one with multiple interpretations. In addition we considered "laughed words," words appearing in the transcripts as *[laughter-yes]* and so on.

For these features, the tables omit values for time until self end, because of a circularity: for example, if a system knows that there will be filler end at a certain time, using that information to infer that a filler start will be more likely a half-second earlier would be circular.

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

We also considered back-channels. In general a person 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 interlocutor is probably likely to continue speaking and perhaps become more fluent and contentful. For simplicity we identified back-channels as tokens labeled *uh-huh* or *um-hum* in the transcription (Hamaker et al., 1998).

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 words might occur soon. Here we defined low pitch regions strictly, as regions where the pitch was consistently below the 20^{th} percentile for that speaker over at least 200 milliseconds. Noteworthy were the good results obtained with time until own next low pitch region. Just before such regions common words included *um-hum, uh-huh*, and *oh*, and relatively uncommon ones included *think, have*, and *don't*; perhaps because such words are common when the speaker is in a fluent patch and about to introduce content, and thus is unlikely to solicit a back-channel soon. Even better performance was obtained with time until own low-pitch region *end*, largely because this gives in addition information about whether a word is overlapping low pitch regions, and it is known that words vary in their affinity for co-occurrence with low pitch regions (Ward, 1999). It is worth noting that this feature is different from all others in this paper in that it uses information that is not purely contextual, but overlapping the word whose identity it is being used to infer.

4.4. Discussion

For practical purposes, the important outcome of this section is the discovery of which features are most informative. Below we explore models which combine these best features. However it is also worth examining some patterns which appear in Tables 1 and 2.

First, features of the behavior of the interlocutor are often as informative as those relating to the behavior of the speaker himself. This probably reflects the highly interactive or "co-constructed" nature of these dialogs.

Second, the time-until features are generally more informative than the time-since features. Perhaps this is because speakers choose words more to indicate what's coming up than they do as an effect of what's happened in the past. Lexical choice obviously depends on many factors, but it seems that our original inspiration, that words reflect cognitive state, which in turn varies with time since landmark events, is probably less important than the fact that words reflect upcoming communicative intentions, especially turn-taking intentions.

Third, some specific features are vastly more useful than others. Those relating to turn-taking are clearly the most important. More surprisingly, fillers are more informative than fragments, although *a priori* both would seem to be equally good indicators of disfluent states. Perhaps the difference is that fragments often lack *communicative* significance, whereas fillers are often other-oriented (Clark and Fox Tree, 2002). If so, again the communicative functions are playing a larger role than the cognitive states.

Interestingly, the most valuable features do not require word hypothesis, which makes them potentially reliable and suitable for use in recognizers operating with noisy inputs. These include obviously utterance boundaries and low pitch regions, but also fillers, backchannels, and laughter, as these are all prosodically and phonetically distinctive (Shriberg, 2001; Ward, 2006; Truong and van Leeuwen, 2007).

5. Prosodic Features

Given the value found for the low-pitch features, we decided to explore the potential of prosody more directly, looking not only at time since a specific prosodic event, but at various properties of the prosodic context immediately prior to a word.

Before presenting our features and results, 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, is the hidden-event model previously discussed (Stolcke et al., 1999), which did use prosodic features related to cognitive factors. In this work, somewhat surprisingly, only durational factors were found to be of value.

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 the performance benefits reported in this section to those of the work cited above.

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

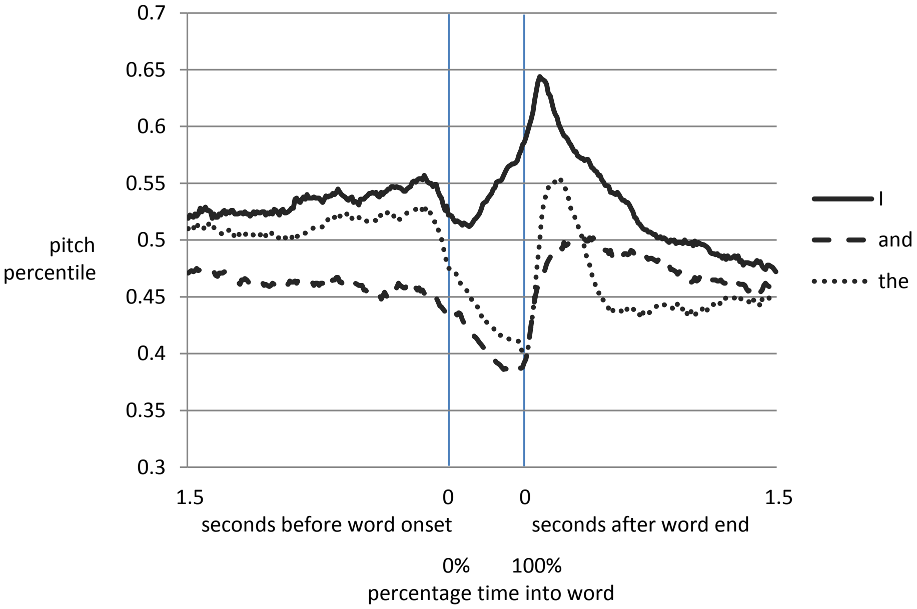

Figure 4: Average pitch in the vicinity of words in Switchboard. Pitch in percentile for that speaker on the y-axis, time on the x-axis: linearly for the 1.5 seconds before the word, scaled within the word to align onsets and endings, and then linearly for the 1.5 seconds after the word.

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 size for computing each feature and then the best values for the k weights. (It is interesting to note that the optimal windows for the prosodic features turned out to be surprisingly short, perhaps in part because the cognitive states of interest are often quite short and their effects immediate, and perhaps in part because they are capturing more lexical stress information than we had expected.) To keep things simple, we only ever used one feature of each type, covering only one small context window, although more benefit could probably be had by using more information from the prosodic contexts of word occurrences, as suggested by Figure 4.

5.1. 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, 2009). 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. As there were only

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 3: Characteristic and Uncharacteristic Words in Different Speaking-Rate Contexts

Conditioning on Local …	Perplexity Reduction	Weight (k)
… speaking rate (from previous word duration)	3.20 %	.99
” (acoustic measure, over previous 325 ms.)	0.95 %	.55
” (mrate, over previous 450 ms.)	0.54 %	.75
… volume (over previous 50 ms.)	2.49 %	.49
… pitch height (over previous 150 ms.)	1.98 %	.60
… pitch range (over previous 225 ms.)	1.60 %	.55

Table 4: Perplexity Improvements on Switchboard, as above.

four buckets, sparseness was less of an issue, so each word was modeled individually; that is, for the prosodic features we did not use an out-of-shortlist class.

We then calculated which words tended to occur in which contexts: Table 3 shows the most characteristic and uncharacteristic, as measured by the scaling factor S. 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, 3.20 % as seen in Table 4. 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, we used the sum of the absolute values of the differences in energy between adjacent 10 ms frames, normalized by the difference between the average speaking volume and the average silence volume, as a very rough approximation to syllable rate. We computed this, over regions of size 325 ms immediately previous to the onset of the word to predict, and 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). We similarly tried a standard better rate estimator, mrate (Morgan and Fosler-Lussier, 1998). In both cases the results were not as good as those obtained using the hand-labeled durations, suggesting that the quality of the rate estimates is important.

5.2. 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 Maximization algorithm to find the mean volumes of the silence regions and of the 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 50 ms. 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 seen in Ward and Vega (2009). As always, the interpretation of such tendencies is not clear-cut. Some 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 perhaps reflect a cognitive state or perhaps a communicative strategy of preceding important words with a quiet lead-in, to give them more impact.

5.3. Pitch Height and Pitch Range

We considered pitch height, as a possible indication of involvement and local dominance, and mindful of evidence that sentence-level contours can affect human lexical access (Braun et al., 2011). Specifically we used the median pitch over the 150 ms. 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, and thus speaker-normalized.

	Weights (k)				
	Two-Feature	Time-Since	Time-Until	Hand-Labeled Combinations	
Feature	Combination	Combination	Combination	Time-Since	Time-Until
time into utterance	.40	.35		.35	
time since interlocutor's end	.45	.40		.40	
time since own low pitch region		.50		.50	
time since other's low pitch region		.20		.20	
time until utterance end			.65		.65
time until interlocutor's start			.55		.55
time until low-pitch start			.05		.00
time until low-pitch end			.60		.06
speaking rate (from labeled duration)				.90	.90
speaking rate (over previous 325 ms.)		.50	.50		
volume (over previous 50 ms.)		.45	.45	.40	.40
pitch height (over previous 150 ms.)		.25	.25	.25	.25
pitch range (over previous 225 ms.)		.10	.15	.05	.05
Benefit (perplexity reduction)	0.7 %	4.6 %	6.8 %	6.3 %	8.4 %

Table 5: Perplexity Reductions for some Combined Models for Switchboard.

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 emotionally colored words relating to belief states such as *admit, surprised, kidding*, and *incredible.*

Pitch range was included as a feature possibly indicative of interest and emotion. Reliability being an issue, we discarded the top and bottom pitchpoints in each 225 ms region and used the ratio between the second-highest and second-lowest pitch points as our measure of range. This value was then compared to the maximum range value seen in a large sample of regions across the entire track, and thus was speaker-normalized. 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.

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.*

6. Combined Models and Use in Speech Recognition

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.

6.1. Combined Models

Wanting to see how well the features would perform together, we took a few subsets of the best-performing features and combined them. Although the features are clearly not independent, we first tried treating them as such. Specifically, we combined all the scaling factors by simply multiplying all of their contributions (Equation 4), that is, by using them all as simultaneous tweaks on the trigram probabilities. A preliminary examination of why some words received penalties rather than benefits (Ward et al., 2010b) revealed that words after pauses were often handled poorly, because of multiple redundant penalties from the various prosodic features. We therefore set the combined models to not apply prosodic tweaks for words at the start of utterances.

We then found good k values for these combined models by hill-climbing. The slow pace of this process was the reason for only using at most 8 features. The resulting models have about 156,000 parameters (24 time slices x 4 temporal features x 1000 words plus 4 categories x 4 prosodic features x 5000 words).

Table 5 shows the results. The Time-Since Combination illustrates what could be obtained for online applications, the Time-Until Combination for offline applications, and the Hand-Labeled Combination indicates what could be approached by using a better rate estimator.

For the very best model, using future features and hand-labeled data, over the whole test set, 63 % of the words had their probability estimates improved by this model. The probability of such an improvement being obtained by chance is less than 0.000001 by the binomial distribution, where the null hypothesis is that our augmentations are improving or degrading the model for each test word randomly. We examined in detail the 10 cases where the contributions of our model most hurt the estimates, relative to the trigram model; most appeared to be due to patterns of behavior that were unusual in this corpus, such as cutting off the other speaker, starting utterances with *uh-huh*, and pausing and sighing at unusual places. This increases our confidence that this model is indeed correctly representing the typical patterns of behavior.

6.2. Results For a Second Corpus, Verbmobil

We also examined a second corpus, the German Verbmobil-II corpus (Jekat et al., 1997). Although also spontaneous speech, in this corpus the speakers' goals were more controlled, limiting the semantic and pragmatic variation seen and the complexity of turn-taking.

This corpus contains 15K 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 46.309.

Due to corpus differences we only examined a few features. In this corpus we identified fillers as tokens labeled *äh*, *ähm*, and *hm* (Batliner et al., 1995). For the temporal features, words outside the 600 most frequent were modeled as a class. The prosodic features were normalized per dialog side, as before, by using the information in all utterances from a speaker in a dialog to order to determine that speaker's typical volume, pitch height, etc.

		Perplexity	Word Error
Feature	k	Benefit	Rate Benefit
time since start of utterance	.10	0.21 %	
time until end of utterance	.20	0.26 %	
time since filler start	.20	0.21 %	
time since filler end	.10	0.08 %	
time until filler start	.60	0.15 %	
speaking rate (over the previous 50 ms.)	.35	−0.03 %	0.0 %
volume (over the previous 200 ms.)	.55	1.69 %	1.0 %
pitch height (over the previous 125 ms.)	.60	1.61 %	0.0 %
pitch range (over the previous 50 ms.)	.45	0.69 %	0.6 %

Table 6: Value for Verbmobil of Each Feature in Isolation

	Weights (k)
speaking rate (over previous 50 necessarily ms.)	.30
volume (over previous 200 ms.)	.35
pitch height (over previous 125 ms.)	.45
pitch range (over previous 50 ms.)	.15
Perplexity Benefit	3.17 %
Word Error Rate Benefit	0.3 %

Table 7: Value for Verbmobil of Combinations of Features

The window sizes for the prosodic features, as seen in the table, were chosen so as to improve performance in the combined model, by reducing their mutual redundancy. Specifically, we used a greedy algorithm, first finding 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 initial search was done with all k values fixed at 0.3.

The results for each feature in isolation are shown in the left columns of Table 6. All perplexity benefits are lower than for Switchboard. This might relate to the fact that the window widths were not retuned, to the less interactive nature of these dialogs, and to differences between German and English in the forms and functions of prosody.

We then built a combined model, as before. The results are seen in Table 7.

6.3. Use in a Speech Recognizer

Although the perplexity improvements obtained were not large enough to be clearly indicative of a benefit for speech recognition (Chen et al., 1998), we went ahead and tried anyway. Not having good acoustic models for English at hand, we tried it on the Verbmobil corpus. The engine used was 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 the prosodic information and combine it with the trigram score whenever the language model was called for a word hypothesis. We tried this both with and without normalization, thinking that normalization might be necessary to ensure that the scores were comparable across the various timepoints when the language model was consulted. However it turned out that the recognition rate was no higher with normalization, despite the enormous computational cost. 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.

As seen in Figures 6 and 7, reductions in word error were indeed obtained for some features and for the combination, with the best case showing an improvement of 0.3 % (1 % relative), significantly outperforming the baseline (matched pairs t-test, on sentence-segment word error rates, $p = 0.023$). The combined model's benefit, however, was lower and not significantly better than the baseline. This suggests, as previous research has also, that setting weights to maximize the perplexity benefit is not necessarily the best way to maximize recognizer performance.

7. 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 up to 8.4 % on Switchboard and 3.2 % on Verbmobil II, both relative to trigram baselines. We have also shown that incorporating these features can reduce a recognizer's word error rate.

These findings suggest that our four principles of Section 2 do indeed describe a promising approach to better language modeling for dialog. This may be true not only for dialog, but also for monologue, as most of the most powerful features found do not use the interlocutor's speech.

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, as suggested by Figure 4. 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 jointly model the features, rather than treating them as independent, perhaps by using a few principal components as a concise characterization of the relevant state of a dialog participant at any moment. In addition, the modeling method should probably be replaced by one that can be combined with n-gram information in better ways, perhaps by using probabilities instead of scaling factors (Section 3), or, more adventurously, by integrating non-lexical predictors into a more general framework (Bengio et al., 2003; Xu and Jelinek, 2007).

Beyond the potential value 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.

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., Kiessling, A., Burger, S., Noeth, E., 1995. Filled pauses in spontaneous speech. Tech. Rep. 88, Verbmobil Project, also in the proceedings of the International Congress of the Phonetics Sciences, 1995.

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., 2009. Variability in word duration as a function of probability, speech style and prosody. Language and Speech 52, 391–413.

Braun, B., Dainora, A., Ernestus, M., 2011. An unfamiliar intonation contour slows down online speech comprehension. Language and Cognitive Processes 26, 350–375.

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.

Clark, H. H., Fox Tree, J. E., 2002. Using *uh* and *um* in spontaneous dialog. Cognition 84, 73–111.

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

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: Conference on Human Language Technologies.

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. 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: Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics.

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., 2001. To 'errr' is human: Ecology and acoustics of speech disfluencies. Journal of the International Phonetic Association 31, 153–169.

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.

Truong, K. P., van Leeuwen, D. A., 2007. Automatic discrimination between laughter and speech. Speech Communication 49, 144–158.

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. G., 1999. Low-pitch regions as dialog signals? Evidence from dialog-act and lexical correlates in natural conversation. In: Swerts, M., Terken, J. (Eds.), ESCA Workshop on Dialogue and Prosody. pp. 83–88.

Ward, N. G., 2006. Non-lexical conversational sounds in American English. Pragmatics and Cognition 14, 113–184.

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

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

Ward, N. G., Vega, A., 2008. Modeling the effects on time-into-utterance on word probabilities. In: Interspeech. pp. 1606–1609.

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-22, University of Texas at El Paso, Department of Computer Science.

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.