

Using Non-Lexical Context to Improve a Language Model for Dialog

Nigel G. Ward, Alejandro Vega

*Computer Science, University of Texas at El Paso
500 West University Avenue, El Paso, Texas 79968 USA*

Abstract

If we can model the cognitive and communicative processes underlying speech, we should be able to better predict what speakers will do, and thus improve language models. This paper presents an initial exploration of this idea. In the Switchboard corpus, we find that word probabilities vary with various non-lexical indicators of cognitive and communicative states, including local volume, local speaking rate and other prosodic features, and also time since start of utterance and since since other reference events. Conditioning word probabilities on 8 such features improved word predictions, reducing the perplexity by 4.4% relative to a trigram baseline.

Key words: dialog state, shallow cognitive state, dialog dynamics, interlocutor behavior, prosody, prediction, word probabilities, perplexity, Switchboard corpus, time, temporal

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), and better predictions 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). One such prediction problem is language modeling: predicting the speaker's next word, given the previous words and other prior context. Having

[☆]We thank Shreyas Karkhedkar for the initial statistics and analysis, Nisha Kiran for the initial affective analysis and modifications to HTK, and Shubhra Datta for incorporating the VoxForge acoustic models; David Novick, Olac Fuentes and several Interspeech and National Science Foundation reviewers for comments; and Joe Picone, Andreas Stolcke, James C. Pennebaker and Margaret Bradley and Peter Lang for making available the Switchboard labeling, the SRILM toolkit, the LWIC dictionaries, and the ANEW data, respectively. This work was supported in part by NSF Grant IIS-0415150 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)

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

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 an assumption that does not really hold for spoken dialog: that the only thing that matters is the lexical context, with other information, such as durations, timing, pitch, and detailed phonetics, being seen as relevant only to the acoustic model. It is probably not coincidental that, despite substantial recent progress, speech recognizer performance is still weak for spontaneous speech in general, and dialog in particular.

In recent years this issue has been addressed by a number of studies. From a functional perspective, most utterances in dialog are there to accomplish something. In task-oriented dialogs the relevant functions can be identified with some precision, and this can be predicted and used in turn to predict upcoming words (Gruenstein et al., 2005; Qu and Chai, 2007). However these methods are limited to dialogs in domains with well-understood semantics, and so there has also been interest in using domain-independent functions, notably dialog acts, such as asking a question, making a statement, or giving back-channel feedback. Jurafsky and colleagues have shown that it is possible to predict the dialog act of an utterance (to some extent) from the interlocutor’s previous dialog act and from the prosody of the utterance; and that the choice of words in an utterance depends in turn (to some extent) on the dialog act of that utterance. This enabled them to build better language models, but the performance increments obtained were disappointing (Jurafsky et al., 1998; Shriberg et al., 1998). An attempt to model the effects of what we might call “sub-acts,” namely the given and new parts of utterances, also gave disappointing results (Ma et al., 2000).

A more direct speaker-interlocutor dependency was found at the “sub-utterance” level: the immediately preceding word of the interlocutor can help predict the next word of the speaker (Ji and Bilmes, 2004), and modeling this gives a substantial perplexity improvement (mostly, it seems, due to conversational routines and semantic priming effects). This work illustrates the potential value of using the interlocutor’s actions as a basis for prediction.

Another interesting series of studies has examined the predictive value of disfluencies (Stolcke et al., 1999). Stolcke and Shriberg showed how these can be represented as “hidden events,” which function, for language modeling purposes, as additional words in the word sequence — a formulation that fits easily into the n-gram framework — and showed that the locations of these hidden events can be inferred in part from the local prosody. Although the quantitative results were disappointing, this work was the first to use a cognition-related performance effect in language modeling.

One question unresolved by the Shriberg-Stolcke studies is whether prosody actually has much value for improving language modeling in dialog. Other uses of prosody in language modeling have seen success (Huang and Renals, 2007) (for broadcast speech using prosody as an indicator of syntactic patterns and lexical item, and for lexically-linked prosodic patterns in meeting recordings), but the value of prosody in its role as an indicator of cognitive states and communicative functions has remained unproven.

Wondering whether the insights underlying these approaches are valid or off-target, we experimented with humans acting as language models, having them hear or view various

types and amounts of context and measuring their ability to predict the next word said (Ward and Walker, 2009). We found that the potential of incorporating such dialog factors exceeded that of just extracting more information from the lexical context. Thus we believe that the mixed results of previous research are due not to errors in the basic conception, but simply to the inability of the techniques so far proposed to accurately model the ways in which word probabilities depend on cognitive and interactive factors.

This paper explores new ways to model the effects of such factors, by conditioning word probabilities on directly observed non-lexical context features. Section 2 presents our strategy of using moment-by-moment variations in dialog state and illustrates the potential by showing how the probabilities of words vary with time into utterance. Section 3 presents our model for representing such tendencies and shows how this information can be combined with that given by a standard n-gram model. Section 4 shows that this reduces perplexity and analyzes the sources of the benefit. Section 5 explores the value of additional reference events, such as time since the end of an utterance by the interlocutor, and Section 6 presents the features with the largest pay-off, those of local prosodic context. Section 7 describes a model combining eight such factors which gives an overall perplexity decrease of 4.4%. Finally Sections 8 and 9 note implications and directions for future work.

2. Time, States, and Events in Dialog

2.1. *Dialog as a Process in Time*

Our modeling strategy is inspired by a theme central to many recent psycholinguistic studies of spoken dialog: the fact that it is a process in time (Clark, 2002, 1996). 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 (variation in word lengths and pause lengths, etc.) 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 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 we see 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 possible to predict, to some extent, what words are likely to occur, not necessarily specific words, but types of words. 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

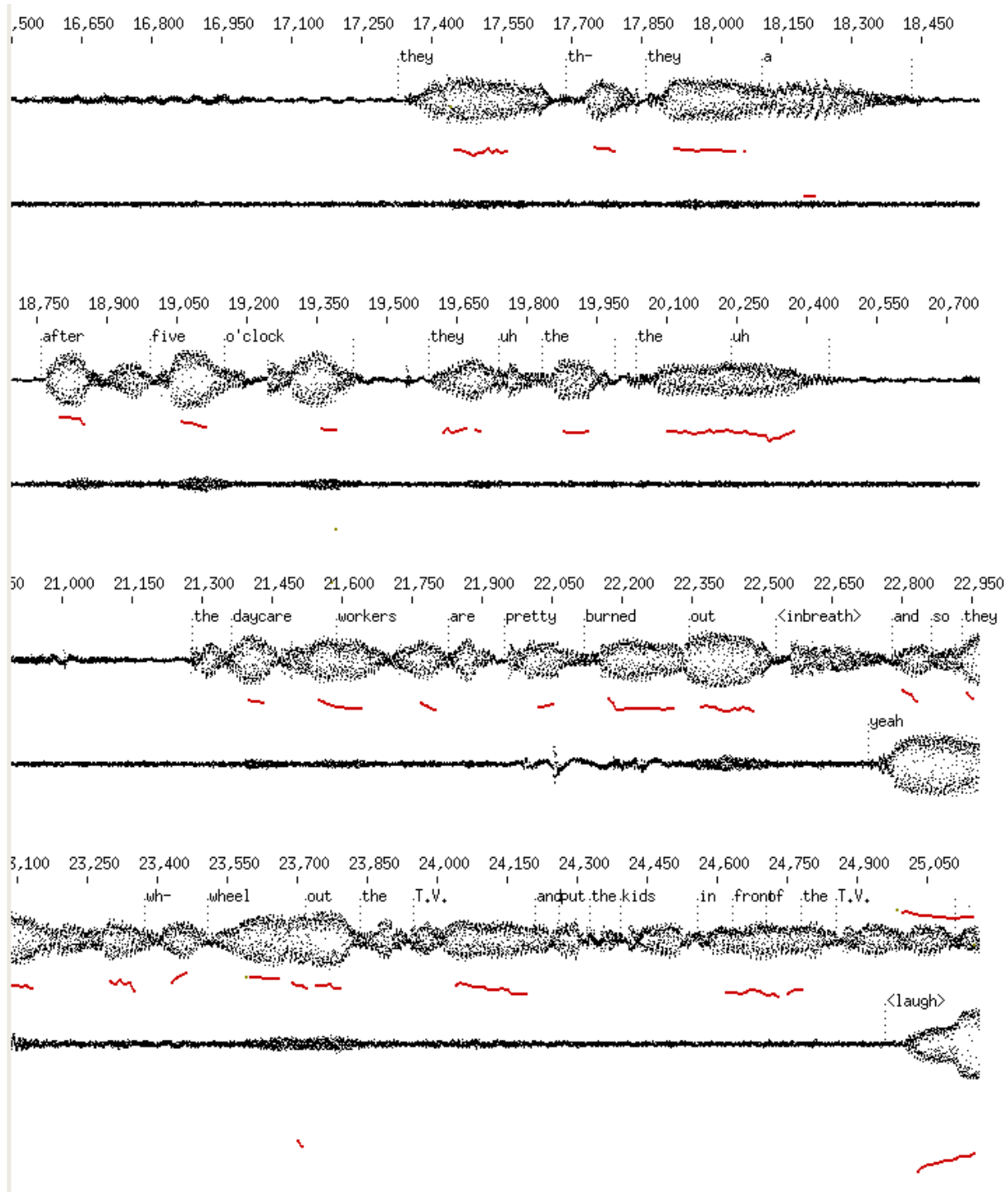


Figure 1: Conversation Fragment. Each of the four strips includes a timeline and two rows, one per speaker. Each row includes a transcription, the signal and the pitch. The second speaker's pitch range is so low that his pitch contours appear far below his signal. This fragment occurred after some talk about television-watching habits and effects on children. (Audio for this clip is available at <http://cs.utep.edu/nigel/abstracts/marc01-excerpt.au>.)

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.

This example illustrates how we may be able to gain leverage for language modeling. The idea can 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 prosody, and
4. this state is somewhat predictive of what word the speaker will say next.

The first principle indicates our intention to model state as it changes moment-by-moment, in contrast to previous models which handle only states that change rarely, for example only at dialog-act boundaries or only in the vicinity of hidden events. The second principle describes the need to represent many facets of dialog state, reflecting the fact that speaker’s minds simultaneously engage in multiple cognitive processes while simultaneously reacting to multiple dimensions of information from the interlocutor.

Taken together, the principles describe the ultimate goal of this line of research; to be 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. Although attributing states to a speaker is unavoidably inexact, it is not mind reading; as previous work has shown, useful information is present not only in the words spoken, but also in the speaker’s pitch, energy, timing, and pronunciation, as well as in the interlocutor’s behaviors.

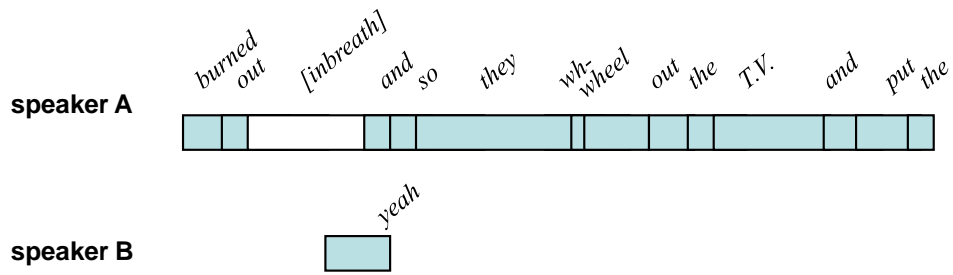
These principles guide our exploration of features and the construction of models in this paper.

2.2. Empirical Interlude: Word Probabilities Vary with Time-into-Utterance

Having presented our perspective and strategy, we now go bottom-up, looking in detail at how word probabilities vary over time, to see what is required of a model. In particular, we examine how word probabilities vary with time-into-utterance, expanding on an earlier study (Ward and Vega, 2008). We chose to look at time-into-utterance thinking that over the course of a typical utterance 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 — 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



A's states and processes

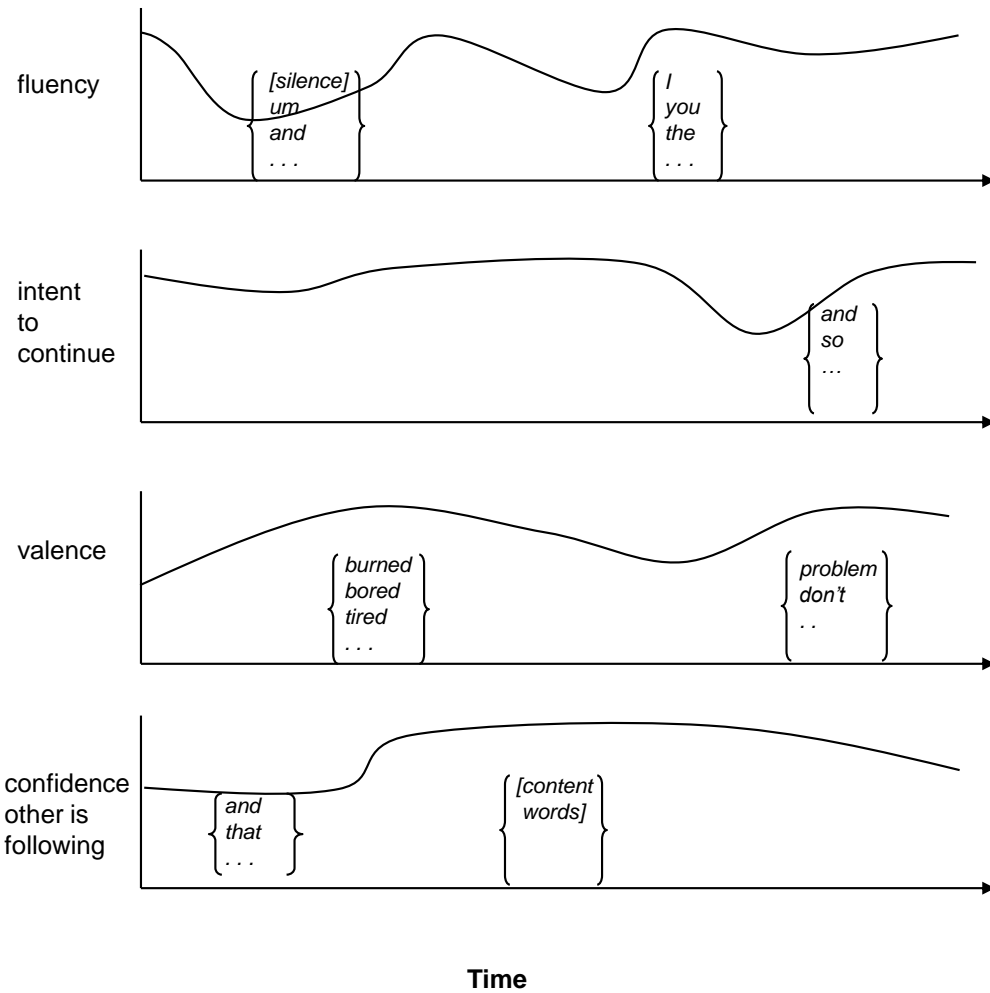


Figure 2: A fanciful image of the aspects of the cognitive state of the speaker in Figure 1 from 22 to 24.5 seconds. Each curve indicates the varying intensity over time of a cognitive state, process, or need. The words in brackets suggest some of words more likely to be produced while in that state.

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.

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:

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

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:

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

Figure 3 illustrates how this ratio can vary over time. The variations suggest that conditioning on time-into-utterance may indeed lead to improved probability estimates.

2.3. Meaningful States and Shallow States

Some of these probability variations are easy to relate to cognitive processes. 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. This may ultimately be due to specific neural pathways; indeed, an imaging study using fMRI has shown “self-referential processing” to be “functionally dissociable from other forms of semantic processing within the human brain” (Kelley et al., 2002).

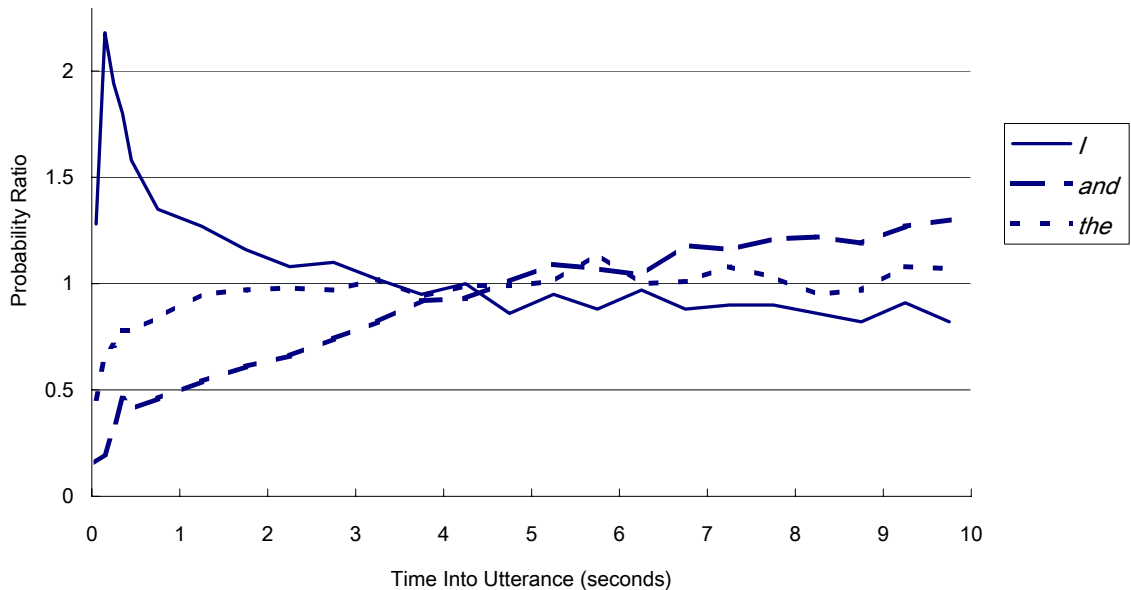


Figure 3: Ratio of Frequency in Specific Time Buckets to Overall Frequency (R) vs. Time-Into-Utterance for the Three Most Common Words. Words at utterance start (time-into-utterance = 0.0) are excluded. The rightmost points represent the range 9.5 seconds and up.

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. 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. Modeling

This section explains how we use such non-lexical context to improve prediction of the next word, continuing to use time-into-utterance to illustrate. As befits our aim, of exploring the possibilities rather than trying to build the ultimate model, the techniques we use are simplistic, with many obvious improvements left untried.

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 .

3.1. Combination with N -grams

The analysis above suggests that time-into-utterance can provide useful information. Clearly, though, this cannot replace the information provided by lexical dependencies. This section accordingly presents a way to use non-lexical context information to improve an existing language model, continuing to use time-into-utterance to illustrate. While the methods described in this section could be used with any language model, we illustrate their use with a trigram model, specifically the SRILM implementation of a backoff model.

This combination of models is a special case of the problem of language model adaptation (Kneser et al., 1997; Bellegarda, 2004), for which many techniques are known. As our aim here is not to determine the best adaptation technique, but merely to determine whether temporal information has value at all, we use only simple methods whose behavior is easy to analyze.

Our first attempt to use time-into-utterance information combined it with the backoff model by linear interpolation, for each word generating a probability estimate using a simple

weighted average of the bucket probability (Equation 1) and the backoff 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 backoff 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 backoff probability by a scaling factor S to reflect this. This gives the “bucket-scaled” backoff probabilities:

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

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 in those 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 we 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 has no actual effect.)

The second complication in the computation of S arises from the fact that the time-based estimate and the backoff 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 backoff probabilities. This we do by raising R to a constant power k less than 1, where a suitable value for k is determined empirically.

Thus,

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

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

backoff probabilities for all the words in the corpus are computed, and the bucket-scaled backoff 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: they must be calculated for each word in the vocabulary. This normalization step, needed for the sake of fair perplexity calculations, makes the amount of computation non-trivial. However this step is not needed for speech recognition.

3.2. Implementation

As our baseline model we used a simple order 3 (trigram) backoff 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).

As a special case, bucket-based scaling is 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 not used for training either. 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.

3.3. Training and Tuning

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. All tokens were converted to lower case.

One meta-parameter is the length-of-pause to use to determine when a new utterance starts — that is, the amount of time that a speaker needs to remain silent in order to end the previous utterance and start a new one. Defining utterances is known to be tricky in general (Traum and Heeman, 1997; Deshmukh et al., 1998). The trade-off for language modeling is that a shorter pause length will, *a priori*, give a large number of short utterances (bad for trigrams) while pushing more words into the early buckets, while a longer pause length will give a smaller set of longer utterances (better for trigrams unless the pause is so long that the pre-pause words are not informative) and push more words into later buckets. We found that the best pause value, both for the baseline trigram model and for the model augmented with time-into-utterance information was 1.2 seconds, although small variations in this value only slightly affected the performance. This value was used in all further investigations. (Parenthetically, there may not be a single best pause value, as pauses of different lengths may tell us different things (Goldman-Eisler, 1967). For example, the initial propensity to use *I* is much weaker if utterances are chosen as talk spurts set off by 2 second

pauses than 1 second pauses; perhaps because the longer planning time makes it easier to retrieve other words or think less egocentrically. For simplicity, however, we use here a single fixed pause value.)

Another set of meta-parameters specifies the widths for the buckets. In general we used the 24 buckets seen in Table 2: namely 5, each 0.1 seconds in width, from 0 to 0.5; 18, each 0.5 seconds in width, from 0.5 to 9.5; and one from 9.5 seconds out to infinity (although a trial with wider buckets to combat data sparsity, notably before 0.5s and after 7.0s, gave a slight performance benefit, of 0.024 points.)

The third meta-parameter was k , which specifies the importance given to the temporal information relative to the n-grams. Best performance was obtained with a value of 0.3.

4. Initial Results and Analysis

Following standard practice, we evaluate language models using perplexity, a measure of the accuracy of predictions. A model assigns probabilities to all words, but its success is judged by its ability to assign high probabilities to the words that actually turn out to occur. This section reports the results of using time-into-utterance information and analyzes the sources of the benefit. It then shows how to use this information in a speech recognizer and briefly considers the use of word classes.

4.1. Initial Perplexity Results

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 ignored sentence-end tags: these also were not predicted.

The results are seen in Table 3. The perplexity was lower for the normalized combined model, indicating that the time-based probabilities are improving the model.

	perplexity
Standard, $P_{backoff}$	107.766
with time-into-utterance, P_n	107.412

Table 3: Initial Evaluation Results

4.2. Patterns of Benefit and Harm

Table 4 illustrates how these computations work. For this example the benefit of the model is strongest for the word *either*, which is common 0.2 to 2 seconds into utterances, and the word *said*, which is common 2 to 3 seconds into utterances. We examined such effects generally across the tuning data, looking for patterns in the ways that it helped and hurt prediction quality.

word	start	bucket	R	S	$P_{backoff}$	P_{bs}	P_n	benefit
well	0.00	–	–	–	.056	.056	.056	—
I	0.15	0.1–0.2s	2.18	1.26	.201	.254	.221	+.042
hadn’t	0.25	0.2–0.3s	1.08	1.00	.001	.001	.001	–.049
either	0.55	0.5–1.0s	2.65	1.34	.005	.007	.006	+.102
we	1.39	1.0–1.5s	0.98	1.00	.004	.004	.004	+.000
hadn’t	1.51	1.5–2.0s	0.29	1.00	.001	.001	.001	–.001
you	1.88	1.5–2.0s	0.95	0.99	.007	.007	.007	–.016
know	1.97	1.5–2.0s	0.82	0.94	.454	.427	.438	–.015
like	2.19	2.0–2.5s	1.11	1.03	.018	.018	.018	+.014
I	2.41	2.0–2.5s	1.08	1.02	.086	.088	.088	+.009
said	2.52	2.5–3.0s	1.27	1.06	.278	.295	.291	+.086
we	2.66	2.5–3.0s	0.93	0.98	.023	.023	.023	–.008

Table 4: Example of the Computations of P_n over a Fragment of an Utterance. The “benefit” is the log of the ratio of P_n to $P_{backoff}$, that is, the difference in their logprobs.

Although generally of value, adding this information had significant negative impact for one of the test dialogs. Upon listening, it turned out the speakers were familiar with each other, and in this respect the dialog was unlike the training data.

In general, utterances that seemed typical of the casual small-talk genre dominating Switchboard were often scored higher, and those less typical were often scored lower. For example, the estimates were hurt for every word in the fluent, grammatical and swift utterance *he does that every year*, especially for the word *every* occurring at 0.48 seconds in, since in Switchboard *every* more typically occurs late in utterances.

Sometimes a word is close to the start, both temporally and in some cognitive or communicative sense, and time-based probability often improved the estimates for such words. For example, an occurrence of *really* at 0.19 seconds into an utterance (after *oh*) was ranked more probable thanks to the time-based model; other words often significantly boosted in this region include *yeah* and *okay*. In a sense, time-based modeling is here capturing a sort of long-distance dependency, allowing the probability estimate for the word *really* to be affected by its proximity to the utterance start. The intervening word, although not skipped, contributes relatively less to the effective context since it is short, thereby enabling *really* to fall in an early bucket.

Sometimes there are words that are much better modeled by trigrams, the word *know* in Table 4 being an example. Here the scaling factor decreases the probability because the unigram *know* is uncommon around 2 seconds in; although in fact the bigram *you know* is relatively common around this time. To alleviate such problems we could base the scaling factor not only on the bucket-based unigrams but also on bucket-based *bigrams*, although sparseness considerations would limit this to the most frequent bigrams.

Sometimes there are words which appear to start a new utterance, in some sense, but

which are not preceded by much silence. These include words that seem to occur more as a response to something said by the interlocutor than as a result of the progress of the speaker’s own cognitive and production processes. For example, in *...sometimes ten to fifteen percent of the an[d]- yeah and and you know the one of the things I remember ...*, the word *yeah* was not preceded by a second of silence. In such cases the bucket-scaling typically decreases the probability of the “restart” word, here *yeah*, decreasing performance. Later we describe a way to combat this problem.

4.3. The Importance of Temporal Information

One might ask whether the information given by time-into-utterance is truly new, or whether it is redundant to that handled by previous models, for example those which simply use more lexical context than trigrams, such as higher-order n-grams, models which use grammatical relations, trigger models, models which support the application of contextual information across wider spans, and models which capture the evolution of cognitive state word-by-word (rather than second-by-second) (Gildea and Hofmann, 1999; Schwenk and Gauvain, 2004; Singh-Miller and Collins, 2007; Ji and Bilmes, 2006).

To determine whether the benefit could really be attributed to the novel idea here, the use of temporal information, we designed a simple experiment. If it were true that the time-of-occurrence of the words didn’t actually matter, then we would see equal or better performance if we conditioned the probabilities not on time-into-utterance, but on word-into-utterance. For example, if the model were merely capturing syntactic regularities (for example, that verbs come in second position or later, *me* comes in third position or later), then a model using word-into-utterance should perform as well or better than a time-based one. Accordingly we built a version of the model where the buckets were based on ordinal position of the word in its utterance. (Computationally, this is equivalent to normalizing with respect to speaking rate, as measured by words per second.) Thus all words in second position fell into one bucket, all words in third position into another, and so on. Words after the 24th position were all cast into one bucket. The R and S values were then computed as above, and the same experiments were run.

We also built a version conditioning probabilities on percent of time into utterance, using ten buckets. This would perform well if the probability variations depended on relative positions in utterances, for example, if there were a tendency that affected all words in the middle of utterances, regardless of whether the utterance is long or short.

Table 5 shows the results. Although conditioning on word-into-utterance also shows a benefit, it is less than that obtained by conditioning on time-into-utterance. Thus the temporal information itself is indeed providing useful extra information.

4.4. Integration in a Speech Recognizer

To explore how to use time-into-utterance information in a speech recognizer, we modified the HTK system (Young et al., 2008). In the lattice rescoring phase, we used the start time for each word hypothesis and looked up the time-based S-value, and then combined it with the probability estimate based on the local context, as explained in Section 3.1. This was done by a simple extension to HLRescore, done by modifying the function *TLatExpand()*

	perplexity	benefit
baseline	107.766	-
time into utterance	107.412	0.354
word into utterance	107.449	0.317
percent into utterance	107.611	0.155

Table 5: Results of Conditioning on Various Measures of Distance into Utterance. The “benefit” is the perplexity decrease relative to the baseline model.

in *HLat.c* to consult not only the standard language model but also the time-based values. These values were taken from simple look-up in an array that had been read in earlier. No normalization was done; that is, we used P_{b_s} instead of P_n , to avoid unnecessary computation. Details of the modifications performed are given in (Kiran and Ward, 2008). This would work also for the other contextual features described below.

In order to properly test whether the new language model actually improves speech recognition would require some additional work, including the creation of acoustic models suitable for Switchboard. Although we have not done this, two preliminary experiments in our laboratory, one with roughly trained acoustic models and one with off-the-shelf (Vox-Forge) acoustic models, obtained decreases in error rates with time-into-utterance information (Kiran and Ward, 2008; Datta, 2009). Also, it is known that decreases in perplexity are generally predictive of decreases in word error rate (Klakow and Peters, 2002), and we have no reason to suspect that this model would be an exception (because there is no reason to think that the time-based contributions would improve the probability estimates only for things that the recognizer would get correct anyway). Thus we think it likely that the perplexity benefits seen are predictive of improved speech recognition.

4.5. Using Word Classes

As seen earlier, some buckets seem to be rich in semantically similar words (Table 2), suggesting that variations in probability with time-into-utterance may be properties of word classes more than of individual words.

To examine this we looked first at two obvious categories: positive and negative emotion words. Using the Affective Norms for English Words, which lists for over 1000 common words the valence, on a 9-point scale (Bradley and Lang, 1999), Nisha Kiran, working in our laboratory, found that words in the first half second of utterances have significantly higher affect on average than words occurring later. We also used LIWC’s positive emotion and negative emotion categories (Pennebaker et al., 2007), and found that the positive emotion words were almost twice as frequent from 0.2 to 0.4 seconds in as elsewhere, whereas negative emotion words are relatively rare until after 0.3 seconds in. It seems likely that these tendencies reflect rhetorical strategies, interpersonal strategies, parameters of memory retrieval processes, and/or cognitive processing constraints.

This suggests that word classes may provide better probability estimates in some cases, especially for less common words. For example, a word like *eighteen* does not occur often

LIWC class	examples	n	benefit
time words	<i>end, until, season</i>	239	0.006
number words	<i>second, thousand</i>	34	0.003
perceptual process words	<i>view, listen, feel</i>	273	0.003
articles	<i>a, an, the</i>	3	0.002
money words	<i>audit, cash, owe</i>	173	0.002

Table 6: Additional perplexity benefits obtained by using class-based S-ratios instead of word-based ratios, for certain classes. n is the number of words in the class in LIWC.

enough in any bucket to provide useful probability estimates (the S-ratios are all 1.0), but when pooled with the counts of other number words, there should be enough data to improve the predictions; that is, using classes may reduce the sparseness problem.

There are many ways to classify words, but for convenience we simply tried out some categories from the LIWC dictionary, including a few that appeared to be relevant from Table 2 (money, time, number, space), a few grammatical categories (verb, preposition, conjunction), the emotional categories (positive, negative), cognitive mechanism (*think, believe* ...) and filler. For each class we used only words of moderate frequency, excluding those among the most frequent 200 words overall, thinking that such words would almost certainly be modeled better by their own specific probabilities, and, as usual, excluding words not among in the top 5000 in the training. The classes that had positive contributions are shown in Table 6; the other classes tried gave nil or negative effects.

As the benefits were not large, we did not further use class-based estimates. However this is probably worth pursuing further: we'd like to consider more classes, soft classes, and classes found automatically by bottom-up clustering, derived perhaps from similarities in the patterns of temporal occurrence relative to various reference events in various corpora. Dimensionality reduction methods are another possible means to the same end (Bengio et al., 2003).

5. Using More Reference Events

This section and the next explore more sources of information, evaluating each new feature using the same models, and the same training, tuning, and test sets.

5.1. Conditioning on Time Since Other's End

Conditioning on time-into-utterance is a useful proxy for conditioning on time since the initiation of the speaking process. It would be perfect if there were always a fixed lag between the cognitive initiation event (whatever that might mean) and the start of vocalization, however this is clearly not always the case. One direction this suggests is 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). However, in line with our strategy of avoiding intricately defined or subjectively labeled events or states,

we went in a simpler direction: we decided to just condition on additional proxies. The first such is the time when the interlocutor ends his turn. Often this will be earlier than the point at which the speaker starts vocalizing, but it may also be later: in Section 4.2 we noted an embedded instance of *yeah* which appeared to be cued by such a turn-end.

We therefore tried conditioning word probabilities on the time since the most recent point at which the interlocutor ended a word and began a significant period of silence. Using pauses of at least 1.2 seconds as delimiters again gave the best performance, and we used a k value of 0.30 again. Used, as before, to enhance a trigram model, this technique gave a fair-sized reduction in perplexity, as seen in Table 7.

Even better, the information in time-since-other’s-end complements 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* is common after self-start, but almost unrelated to other-end.

	perplexity benefit
time into utterance	0.354
time since other’s end	0.341
combined, k retuned	0.661

Table 7: Results for Conditioning with Respect to Two Reference Events, and their Combination

Combining the two sources of information is easy: we tweak the trigram probabilities using both, by adding an additional multiplicative factor to Equation 3. Doing this we found 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 7. 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 relative to the trigram information.

Using times since multiple reference events gives a sort of multi-layered representation of the speaker state at any time. The perplexity benefit seen provides suggestive evidence for the value of modeling dialog behavior as arising from the simultaneous operation of multiple cognitive and communicative processes. While not common in language modeling this matches well with the view of dialog researchers who stress the simultaneous presence of dimensions of emotional, attitudinal, meta-communication and interpersonal communication in parallel to the communication of content (Goffman, 1981; Clark, 1996; Brennan and Hulstén, 1995; Campbell, 2007; Petukhova and Bunt, 2009), and those who point out 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).

5.2. Conditioning on Time Since Other Reference Events

Emboldened by this success, we tried more reference events, choosing events which seemed to relate to cognitive or communicate state, and which could be found in or computed

feature	best weight (k)	benefit
time into utterance	.30	0.354
time since interlocutor’s end	.40	0.364
onset of own last filler	.25	0.104
end of own last filler	.35	0.076
onset of other’s last filler	.25	0.151
end of other’s last filler	.35	0.101
time since own last fragment	.35	0.130
time since other’s last fragment	.40	0.020
time since own back-channel	.35	0.132
time since other’s back-channel	.40	0.149
time since own low-pitch region	.50	0.226
time since other’s low-pitch region	.45	0.218
time since own laughter onset	.50	0.082
time since own laughter end	.80	-0.038
time since other’s laughter onset	.70	0.089
time since other’s laughter end	.55	0.022

Table 8: Perplexity Improvements obtained by conditioning on various reference events.

easily from the audio signal or the transcripts.

In Switchboard many utterances start with a filler word 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*. As seen in Table 8, fair results were seen for time since the end of one’s own last filler.

Word fragments are also common in Switchboard, and may be a good proxy for disfluency events, indicating whether a speaker has his utterance planned out ahead or not. 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*. The benefits were small.

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 events. Fair results were seen.

Because speakers of English frequently invite back-channel feedback from the interlocutor

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

by producing low pitch regions (Ward and Tsukahara, 2000), we thought that the presence of such regions may also provide clues as to what sort of information they might be preparing to produce. Fairly good results were obtained.

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.

6. Using Local Prosodic Information

Another source of information about the cognitive state of a speaker is the prosodic features of his or her recent speech. A number of previous studies have used prosody in language modeling and speech recognition, as surveyed in (Shriberg and Stolcke, 2004) and (Huang and Renals, 2007). Our choices of which 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, not hand-labeled nor inferred to match hand-labeled tags; 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 . The quantitative results appear in the bottom half of Table 10 below.

6.1. 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). 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 classified as 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 9 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. 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. We also tried normalizing with respect to the speaker’s overall rate, such that a word would be considered fast or slow relative 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 of course 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, we obtained a very rough approximation to syllable rate. We computed this, over regions of size 325ms, namely the regions immediately previous to 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, seen in Table 10 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 indicated.

6.2. Volume

We considered volume, as likely to indicate states such as engagement or dominance. Volume was computed over 50ms timespans and was speaker-normalized. Specifically, for each dialog side we took at a large sample of timespans, and used EM 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 timespan immediately preceding the word onset.

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*,

mostly), 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. As always, the interpretation of such tendencies is not clear-cut. On the one hand, they do seem to reveal cognitive states. On the other hand, one could also interpret these patterns of occurrence as reflecting communicative situations: for example, the tendency for expressions of belief to come after low-volume regions may reflect a communicative strategy of preceding important words with a quiet lead-in, to give them more impact. For practical language modeling purposes, the interpretation given doesn't matter: what matters is whether regularities exist, and whether they are non-redundant to those captured by standard language models, as revealed by the perplexity results.

A fair perplexity improvement was obtained, as seen in Table 10. Overall, the words that gave maximum benefit were *yeah, oh, um-hum, uh-huh, well* and *and*, which were more common in the after-silence condition, and *to* and *of*, which were uncommon in this condition. Words in the after-loud condition also contributing strongly, mostly due to words like *to, a, it, have, and of*, which tend to be more common after loud regions.

6.3. Pitch Height

We considered pitch height, as a possible indication of involvement and local dominance. We computed the median pitch over the 150ms immediately preceding the onset of the word to predict. We 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, as seen in Table 10. 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*.

6.4. Pitch Range

Pitch range was included as a possible indication 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. We 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, as seen in Table 10. 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, and nice*, social words such as *work* and *family*, and the shopping words *bought* and *spend*.

feature	best k in isolation	benefit in isolation	k in the combined model
time into utterance	0.30	0.354	0.40
time since other’s end	0.40	0.364	0.25
time since own low pitch region	0.50	0.226	0.25
time since other’s low pitch region	0.45	0.218	0.25
speaking rate (over the previous 325ms.)	0.55	1.136	0.45
volume (over the previous 50ms.)	0.49	2.651	0.45
pitch height (over the previous 150ms.)	0.60	2.046	0.30
pitch range (over the previous 225ms.)	0.55	1.741	0.10
combined	—	4.788	—

Table 10: Perplexity Improvements obtained by tuning weights and combining features. The “benefit” is the perplexity decrease relative to the baseline model; the third column shows this for each feature in isolation with the k value shown in the second column. The rightmost column is the best weight for that feature in the combined model, whose results are shown at the end.

7. A Combined Model

We picked the 8 best-performing features found and built a combined model by multiplying all of their contributions (Equation 4), that is, by using them all as simultaneous tweaks on the trigram probabilities. We found good k values for this combined model by hill-climbing. Table 10 shows the results: the total perplexity reduction obtained was points, which is 4.4%.

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, using optimized k values, gave a perplexity of 102.025, which is a 5.741 benefit: a 5.3% decrease in perplexity.

8. Directions for Future Work

8.1. Improvements to the Model

At this point we have amassed evidence in support of the idea that the four principles of Section 2.1 do indeed describe a promising strategy. To progress further it is probably worth devising a new modelling method: replacing the one developed here, which was convenient for empirical explorations, with one better suited for the full exploitation of the potential of non-lexical predictors. Thus we should find better ways to combine non-lexical information with trigrams, or alternatively, to integrate non-lexical predictors into a more general framework (Bengio et al., 2003; Xu and Jelinek, 2007).

In particular, we need to reduce the number of parameters in the model, which are already numerous, even without promising additions such as features for voicing and other phonetic variations, more features of the interlocutor’s recent behavior, and time *until* event

(negative time) features. One way to do this would be to use word classes, as suggested above. Another way would be to replace estimates based on individual parameters for each bucket with more sophisticated, smoother probability estimators. Another approach would be to develop more concise representations of state: rather than using a multi-dimensional model (e.g. the 8 dimensions for the combined model above), somehow characterizing the relevant state of a dialog participant at any moment more concisely, perhaps in terms of a few principal components. Alternatively we may be able to find equivalence classes of context, perhaps even classes that capture the effects both of lexical and non-lexical context.

8.2. Applications

The importance of better language models for speech recognition is clear. As the sources of information introduced here do not overlap those handled by existing models, we expect the perplexity improvements to be matched by improvements in recognition rates. We also expect the methods to work not only for dialog but for any type of speech where the output is affected by the ongoing cognitive processes of the speaker.

Beyond the performance improvements already demonstrated, models using non-lexical information have the potential to be more robust: to the extent that the patterns of word occurrence they exploit are reflecting fundamental cognitive processes and constraints, such language models are more likely to transfer well across domains and tasks than are, say, pure n-grams, which are known to be brittle (Bellegarda, 2004). Such models may also be useful for predictions that go beyond words, to also 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 is crucial for a full understanding of dialog behavior, which must include those aspects typically missing from written representations.

Thus we would like to create a “dialog model,” like a language model, but going beyond words to also predict other aspects of what a speaker is likely to do (or did do in the test data) in the next moment. This subsumes the language modeling problem and a number of problems in dialog management. These include speech-versus-silence prediction (also known as endpointing) (Ferrer et al., 2003; Raux and Eskenazi, 2009), and smooth turn-taking more generally. They also include adaptation, both of low-level features such as speaking rate and those coding the interpersonal, attitudinal and emotional dimensions of interaction in dialog, as seen in response patterns over short time frames (Ward and Nakagawa, 2004; Acosta and Ward, 2009). These topics have recently received much attention, in part because spoken dialog systems today are awkward and inefficient to use, as a way to make their behavior better conform to human norms of turn-taking and other interpersonal behaviors (Shriberg, 2005; Ward et al., 2005) and there are many relevant findings which may be easier to extend, integrate and exploit if built into a unified predictive model.

8.3. Other Prospects

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 foundation for more general predictive models, and thus a valuable addition to the set of tools for understanding human behavior. Our finding here, that word probabilities vary with shallow cognitive states, may suggest a path to the development of deeper models of the cognitive processes underlying dialog.

Such models are not only of scientific interest. Areas that this line of inquiry may impact include intercultural training, clinical pragmatics, and assessment of interpersonal dynamics (Kiekel et al., 2002; Niederhoffer and Pennebaker, 2002; Pentland, 2008; Jurafsky et al., 2009). More generally, improving communication skills is important for many people, as witnessed by the large number of self-improvement books on how to be a more effective public speaker, negotiator, conversationalist, listener, etc. This reflects the importance of good dialog skills for effectiveness, success, and happiness, not only for individuals but also for society. This line of inquiry may lead to useful findings about the ways that (successful) speakers allocate cognitive effort and respect the cognitive constraints of their interlocutors, potentially showing the way to improved communication for everyone.

9. Conclusion

This paper identified new factors useful for language modeling, directly using the ways in which word probabilities vary with non-lexical context, including local prosodic information and elapsed times since various reference events, and showed that this information can be exploited in a language model, giving a 4.4% perplexity reduction over a baseline trigram model on the Switchboard corpus.

Today most language models treat speech as if it were text, as simply sequences of words. Spoken language is, however, created by human minds and for human minds Fujisaki (2008). This study is a step away from viewing language modeling as a purely symbolic enterprise, and a step towards a view better suited to modeling *spoken* language, based on the recognition that lexical behavior is integrated with other language behaviors, especially prosodic behaviors, and that lexical behavior inextricably reflects the cognitive state of the speaker and the dynamics of the interaction with the interlocutor.

References

- Acosta, J. C., Ward, N. G., 2009. Responding to user emotional state by adding emotional coloring to utterances. In: Interspeech.
- 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.

- 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.
- Bradley, M. M., Lang, P. J., 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Tech. Rep. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.
- Brennan, S. E., Hulstien, 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.
- Clark, H. H., 1996. *Using Language*. Cambridge University Press.
- Clark, H. H., 2002. Speaking in time. *Speech Communication* 36, 5–13.
- Datta, S., 2009. Various test results, UTEP CS ISG internal memo, May 8, 2009.
- Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., Picone, J., 1998. Resegmentation of Switchboard. In: *ICSLP*. pp. 1543–1546.
- 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.
- Gildea, D., Hofmann, T., 1999. Topic-based language models using EM. In: *Eurospeech*.
- 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.
- Goldman-Eisler, F., 1967. Sequential temporal patterns and cognitive processes in speech. *Language and Speech* 10, 122–132.
- 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*.
- Gruenstein, A., Wang, C., Seneff, S., 2005. Context-sensitive statistical language modeling. In: *Interspeech*.
- 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., Boulard, 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.
- 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., 2006. Backoff model training using partially observed data: Application to dialog act tagging. In: *HLT/NAACL*.

- Jurafsky, D., Ranganath, R., McFarland, D., 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In: Proceedings of NAACL HLT.
- Jurafsky, D., Shriberg, E., Fox, B., Curl, T., 1998. Lexical, prosodic, and syntactic cues for dialog acts. In: Association for Computational Linguistics, Workshop on Discourse Relations and Discourse Markers.
- Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., Heatherton, T. F., 2002. Finding the self ? an event-related fMRI study. *Journal of Cognitive Neuroscience* 14, 785–794.
- Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J., Martin, M., 2002. Some promising results of communication-based automatic measures of team cognition. In: Proceedings of the Human Factors and Ergonomic Society. Vol. 46.
- Kiran, N., Ward, N. G., 2008. Testing the value of a time-based language model for speech recognition. Tech. Rep. UTEP-CS-08-29, University of Texas at El Paso, Department of Computer Science.
- Klakow, D., Peters, J., 2002. Testing the correlation of word error rate and perplexity. *Speech Communication* 38, 19–28.
- Kneser, R., Peters, J., Klakow, D., 1997. Language model adaptation using dynamic marginals. In: Proc. Eurospeech. pp. 1971–1974.
- 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.
- Niederhoffer, K. G., Pennebaker, J. W., 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21, 337–360.
- O’Connell, D. C., Kowal, S., 2008. *Communicating with One Another: Toward a Psychology of Spontaneous Spoken Discourse*. Springer.
- Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., Booth, R. J., 2007. The development and psychometric properties of LIWC2007. Tech. rep., LIWC.net (Linguistic Inquiry and Word Count).
- Pentland, A., 2008. *Honest Signals*. MIT Press.
- 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.
- Qu, S., Chai, J. Y., 2007. An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In: NAACL HLT. pp. 284–291.
- 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.
- Schwenk, H., Gauvain, J.-L., 2004. Neural network models for conversational speech recognition. In: Interspeech.
- Sebanz, N., Bekkering, H., Knoblich, G., 2006. Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences* 10, 70–76.
- Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., Van Ess-Dykema, C., 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech* 41, 439–487.
- Shriberg, E., Stolcke, A., 2004. Prosody modeling for automatic speech recognition and understanding. In: *Mathematical Foundations of Speech and Language Processing, IMA Volumes in Mathematics and Its Applications*, Vol. 138. Springer-Verlag, pp. 105–114.
- Shriberg, E. E., 2005. Spontaneous speech: How people really talk, and why engineers should care. In: *Interspeech*. Lisbon.

- Singh-Miller, N., Collins, M., 2007. Trigger-based language modeling using a loss-sensitive perceptron algorithm. In: IEEE ICASSP.
- 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.
- Traum, D., Heeman, P., 1997. Utterance units in spoken dialogue. In: Maier, E., Mast, M., LuperFoy, S. (Eds.), Processing in Spoken Language Systems. Springer-Verlag, pp. 125–140.
- Ward, N., Nakagawa, S., 2004. Automatic user-adaptive speaking rate selection. *International Journal of Speech Technology* 7, 235–238.
- Ward, N., Rivera, A. G., Ward, K., Novick, D. G., 2005. Root causes of lost time and user stress in a simple dialog system. In: Interspeech.
- 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., Vega, A., 2008. Modeling the effects on time-into-utterance on word probabilities. In: Interspeech. pp. 1606–1609.
- Ward, N. G., Walker, B. H., 2009. Estimating the potential of signal and interlocutor-track information for language modeling. In: Interspeech.
- 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.
- Young, S., et al., 2008. The HTK book, from <http://htk.eng.cam.ac.uk/docs/docs.shtml>.