Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?

Elizabeth Shriberg, SRI International, Menlo Park, CA, U.S.A.
Rebecca Bates, Boston University, Boston, MA, U.S.A.
Paul Taylor, University of Edinburgh, Edinburgh, U.K.
Andreas Stolcke, SRI International, Menlo Park, CA, U.S.A.
Daniel Jurafsky, University of Colorado, Boulder, CO, U.S.A.
Klaus Ries, Carnegie Mellon University, Pittsburgh, PA, U.S.A.
Noah Coccaro, University of Colorado, Boulder, CO, U.S.A.
Rachel Martin, Johns Hopkins University, Baltimore, MD, U.S.A.
Marie Meteer, BBN Systems and Technologies, Cambridge, MA, U.S.A.
Carol Van Ess-Dykema, U.S. Department of Defense, Ft. Meade, MD, U.S.A.

Running Head: Prosodic detection of dialog acts

Acknowledgements: We thank the organizers and sponsors of the WS97 Workshop on Innovative Techniques in LVCSR at the Center for Speech and Language Processing at Johns Hopkins University. Additional support was provided by the NSF through grants IRI-9619921 and IRI-9314967. Special thanks go to the Boulder graduate students for the discourse labeling, to Mitch Weintraub of SRI for SNR measurement software, and to Nelson Morgan, Nikki Mirghafori and Eric Fosler at ICSI for making the enrate software available. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies of the funding agencies.

Corresponding Author:

Elizabeth Shriberg
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
Tel: +1-650-859-3798
FAX: +1-650-859-5984

ABSTRACT

Identifying whether an utterance is a statement, question, greeting, and so forth is integral to effective automatic understanding of natural dialog. Little is known, however, about how such dialog acts (DAs) can be automatically classified in truly natural conversation. This study asks whether current approaches, which use mainly word information, could be improved by adding prosodic information.

The study examines over 1000 conversations from the Switchboard corpus. DAs were hand-annotated, and prosodic features (duration, pause, F0, energy and speaking-rate features) were automatically extracted for each DA. In training, decision trees based on these features were inferred; trees were then applied to unseen test data to evaluate performance.

For an all-way classification as well as three subtasks, prosody allowed highly significant classification over chance. Feature-specific analyses further revealed that although canonical features (such as F0 for questions) were important, less obvious features could compensate if canonical features were removed. Finally, in each task, integrating the prosodic model with a DA-specific statistical language model improved performance over that of the language model alone. Results suggest that DAs are redundantly marked in natural conversation, and that a variety of automatically extractable prosodic features could aid dialog processing in speech applications.

INTRODUCTION

Identifying whether an utterance is a statement, question, greeting, and so forth is integral to understanding and producing natural dialog. Human listeners easily discriminate such dialog acts (DAs) in everyday conversation, responding in systematic ways to achieve the mutual goals of the participants (Clark, 1996; Levelt, 1989). Little is known, however, about how to build a fully automatic system that can successfully identify DAs occurring in natural conversation.

At first blush, such a goal may appear misguided, because most current computer dialog systems are designed for human-computer interactions in specific domains. Studying unconstrained human-human dialogs would seem to make the problem more difficult than necessary, since task-oriented dialog (whether human-human or human-computer) is by definition more constrained and hence easier to process. Nevertheless, for many other applications, as well as for basic research in dialog, developing DA classifiers for conversational speech is clearly an important goal. For example, optimal automatic summarization and segmentation of natural conversations (such as meetings or interviews) for archival and retrieval purposes requires not only knowing the string of words spoken, but also who asked questions, who answered them, whether answers were agreements or disagreements, and so forth. Another motivation for speech technology is to improve word recognition. Because dialog is highly conventional, different DAs tend to involve different word patterns or phrases. Knowledge about the likely DA of an utterance could therefore be applied to constrain word hypotheses in a speech recognizer. Modeling of DAs from human-human conversation can also guide the design of better and more natural human-computer interfaces. On the theoretical side, information about properties of natural utterances provide useful comparison data to check against descriptive models based on contrived examples or speech produced under laboratory settings. Development of automatic methods for classifying dialog acts could also be applied to the problem of labeling large databases when hand-annotation is not feasible, thereby providing data to further basic research.

This project is part of a larger project (Jurafsky et al., 1997a; Jurafsky et al., 1997b) investigating approaches for the automatic modeling of DAs in truly natural human-human dialogs. Results from the larger project, as well as related studies on German human-human task-oriented dialog (Mast et al., 1996)

indicate that a primary knowledge source for DA classification is the words (either true words, or output from a speech recognizer). Many DAs can be discriminated to a large extent using a statistical language model that stores separate word probability distributions for each DA. The current paper focuses on exploring another, relatively untapped potential knowledge source for automatic DA classification: prosody. By prosody we mean information about temporal, pitch and energy characteristics of utterances that are independent of the words.

We were interested in prosody for a number of reasons. First, some DAs are inherently ambiguous from word information alone. For example, declarative questions (e.g. "John is here?") have the same word order as statements, and hence are only distinguishable from statements via prosody. Second, in a real application the word accuracy may not be perfect. Indeed, state of the art recognizers are still at over 30% word error rate for large vocabulary conversational speech. Third, there are potential applications for which one may not have available full fledged speech recognition, but rather want to track roughly what is happening in a dialog by some other means. Fourth, an understanding of prosodic properties of different utterance types can improve the naturalness of speech synthesis systems. And finally, it is of basic theoretical interest to descriptive accounts in linguistics, as well as to psycholinguistic theories of sentence processing, to understand how different DAs are signalled prosodically.

Although there has been little work on the problem of automatic DA classification for unconstrained human-human conversations, a number of related studies on prosody and discourse provide a helpful starting point for this work. The meaning and function of intonational contours in discourse has been described within a theoretical framework (Pierrehumbert and Hirschberg, 1990; Grosz and Hirschberg, 1992; Hirschberg and Nakatani, 1996). Prosodic properties of particular dialog acts (e.g., questions and statements) have been described in studies using hand-measured features in read or elicited speech (Vaissière, 1983; Haan et al., 1997a; Haan et al., 1997b; van Heuven et al., 1997). Prosodic features associated with discourse structure have also been described, particularly for human-computer dialogs (Geluykens and Swerts, 1993; Swerts, 1997; Swerts and Ostendorf, 1997). Furthermore there has been increasing interest in studies of dialog-act classification for human-human dialog in a task-oriented setting (Batliner et al., 1993; Mast et al., 1996;

Taylor et al., 1997a; Warnke et al., 1997).

The present study focusses on unconstrained human-human dialog, and more specifically on inherent properties of the utterances themselves. Thus the goal differs from that of the overall project described in Jurafsky et al. (1997a), which aimed to optimize overall classification performance for the naturally-occurring distribution of DAs. In the larger project, classification involved the integration of three knowledge sources: (1) DA-specific LMs, (2) a dialog grammar (a statistical model of the sequencing of DAs in a conversation), and (3) DA-specific prosodic models. Results from preliminary experiments revealed that the modeling was driven largely by priors (encoded as unigram frequencies in the dialog grammar) because of an extreme skew in the distribution of DAs in the corpus. Since here we seek to understand whether prosodic properties of the utterances themselves can be used to predict DAs, we eliminate additional knowledge sources that could confound our results. Analyses are conducted in the "no-priors" domain (all DAs are made equally likely). We also exclude contextual information from the dialog grammar (such as the DA of the previous utterance). In this way, we hope to gain a better understanding of the prosodic properties of the different DAs, which can in turn be applied in building better integrated models for natural speech corpora in general.

Our approach builds on recent methodology that has been applied to conversational speech for a different task (Shriberg et al., 1997). The method involves construction of a large database of automatically extracted acoustic-prosodic features. In training, decision tree classifiers are inferred from the features; the trees are then applied to an unseen set of data to evaluate performance.

The analyses examine tree performance in four DA-classification tasks. We begin with a task involving all-way classification of the DAs in our corpus. We then examine three subtasks found to be problematic for word-based classification: question detection, agreement detection, and the detection of incomplete utterances. For each task, we train classifiers using various subsets of features to gain an understanding of the relative importance of different feature types. In addition we integrate tree models with DA-specific language models to explore the role of prosody when word information is also available.

METHOD

Speech Data

Our data were taken from the Switchboard corpus of human-human telephone conversations on various topics (Godfrey et al., 1992). There were clear advantages to using this corpus, including its size, the availability of N-best word recognition output from a state-of-the-art recognition system, and the representation of many different speakers. The quality of the speech is clearly "natural" in that transcribers asked to rate naturalness used the highest value in the scale for the majority of speakers and conversations.

Dialog Act Labeling

We developed a DA labeling system for Switchboard, taking as a starting point the DAMSL system (Allen and Core, 1997) of DA labeling for task-oriented dialog. We adapted the DAMSL system to allow better coverage for Switchboard, which differs from the task-oriented forms of dialog for which DAMSL was developed. Certain classes in DAMSL were never used, and conversely it was necessary to expand some of the DAMSL classes to provide a variety of labels. The adapted system, "SWBD-DAMSL," is described in detail in Jurafsky et al. (1997).

Since there was a large set of data to label, labeling was done using the transcriptions of the full conversations, but without actually listening to the utterances. (A similar approach was also taken, for similar reasons, in the work of Mast et al. (1996). Labelers were instructed to flag utterances which they felt were ambiguous from text alone; in the present analyses all such flagged utterances are excluded. Labeling from text greatly speeds the process, but causes an inherent bias toward DA labels that are consistent with the words and the discourse context. This means that DAs that are conveyed mainly prosodically have a higher risk of being mislabeled. A clear example is the distinction between simple backchannels, which acknowledge a contribution (e.g. "uh-huh") and explicit agreements (e.g. "that's exactly it"). There is considerable overlap lexically between these two DAs, with emphatic intonation conveying an agreement (e.g., "right" versus "right!"). Emphasis of this sort was not marked by punctuation in the transcriptions, and backchannels were nearly four times as likely in our corpus, so some agreements are likely to be mislabeled as backchannels in our data. In general, because any labeling bias works *against* our hypothesis, i.e., makes it more difficult to detect DAs using prosodic information, the approach was warranted for these analyses.

We expect that results would only improve if listening were included in the labeling procedure.

SWBD-DAMSL defines approximately 60 unique tags, many of which represent orthogonal information about an utterance and hence can be combined. The labelers made use of 220 of these combined tags, which we clustered for our larger project into 42 classes (Jurafsky et al., 1997a). To simplify analyses for the present paper, the 42 classes were further grouped into seven orthogonal main classes, consisting of the frequently-occurring classes plus an "other" class containing DAs each occurring less than 2% of the time.

The groups are shown in Table 1. The full set of DAs is listed in the Appendix, along with actual frequencies. The full list is useful for getting a feel for the heterogeneity of the "other" class.

For the statement classes, independent analyses showed that the two SWBD-DAMSL types of statements, descriptions and opinions, were quite similar in their lexical and their prosodic features, although they did show some differences in their distribution in the discourse which warrants their continued distinction in the labeling system. Since, as explained in the Introduction, we do not use dialog grammar information in this work, there is no reason not to group the two types together for analysis purposes. For the Question category we also extracted from the set the main question types described by Haan et al. (1997a), namely declarative questions, yes-no questions, and wh-questions, as well as miscellaneous question types.

Table 1: Seven grouped dialog act classes

Type	SWBD-DAMSL Tag	Example
Statements		
description	sd	*Me, I'm in the legal department*
view/opinion	sv	*I think it's great*
Questions		
yes/no	qy	*Do you have to have any special training?*
wh	qw	*Well, how old are you?*
declarative	qy^d, qw^d	*So you can afford to get a house?*
open	qo	*How about you?*
Backchannels	b	*Uh-huh*
Incomplete Utts	%	*So, -*
Agreements	aa	*That's exactly it*
Appreciations	ba	*I can imagine*
Other	all other	(see appendix)

Interlabeler reliability was assessed using the Kappa statistic (Siegel and Castellan, Jr., 1988; Carletta, 1996), or the ratio of the proportion of times that raters agree (corrected for chance agreement) to the maximum proportion of times that the rates could agree (corrected for chance agreement). Kappa computed for the rating of the original 42 classes was 0.81, which is considered high for this type of task. *Post hoc* grouping of the ratings using the seven main classes just described yielded a Kappa of 0.85.

Dialog Act Segmentation

For an actual system, DA classification also means finding the boundaries between utterances. The issue of how to automatically segment utterances in conversational speech has gained interest in recent years by a number of researchers. An approach to DA classification that attempts to also find utterance boundaries has been described by Mast et al. (1996). This is clearly the problem that must ultimately be solved. One complication in interpreting results, however, is that turn boundaries, which are available "for free," correlate with segment boundaries; thus at turn boundaries more information is known than at boundaries within a speaker's turn. For this work we did not want to confound the issue of DA classification with DA segmentation, or to treat DAs at turn boudaries better than those not at turn boundaries, thus we used boundaries as marked by labelers as "slash units" according to the LDC annotation guidelines described in Meteer et al. (1995). To keep results of different knowledge sources comparable, these DA boundaries were also made explicit for the speech recognition and language modeling used.

The slash units were marked between words. To estimate the locations of the boundaries in the speech waveforms, a forced alignment of the acoustic training data was merged with linguistically annotated training transcriptions from the LDC. This yielded word and pause times of the training data with respect to the acoustic segmentations. Using these word times along with the linguistic segmentation marks, the start and stop times for linguistic segments were found.

This technique was not perfect, however, for several reasons. One is that many of the words included in the more careful linguistic transcription had been excised from the acoustic training data. Some speech segments were considered not useful for acoustic training and thus has been excluded deliberately. In addition the alignment program was allowed to skip words at the beginning and end of an acoustic segment

if there was insufficient acoustic evidence for the word. This caused problems in the context of highly reduced pronunciations or for low-energy speech, both of which are frequent in Switchboard. If times were available for some words in an utterance, even though the end words were missing times, we noted the available times as well as how many words were missing from the times and if they were at the beginning or end (or both) of the utterance.

Errors in the boundary times for DAs crucially effect the prosodic analyses, since prosodic features are extracted assuming the boundaries are reasonably correct. Incorrect estimates affect the accuracy of global features (e.g., DA duration), and may render local features meaningless (e.g. F0 measured at the supposed end of the utterance). Since features for DAs with known problematic end estimates would be misleading in the prosodic analyses, they were omitted from our training (TRN) and held-out test (HLD) data.

Overall, we were missing 30% of the training utterances because of problems with time boundaries. While the majority of the words in the training data were included (i.e., enough data for acoustic modeling purposes), we were missing up to 45% of some types of utterances, backchannels in particular. While these utterances may not contribute to a significant drop in error rate, they are important for modeling the flow of the conversation. The time boundaries of the DEV test set, however, were carefully handmarked for other purposes, so we were able to use exact values for this test set. It should be noted that this difference in segmentation method makes the DEV test set somewhat mismatched with respect to the training data.

Prosodic Features

The prosodic database included a variety of features that could be computed automatically, without reference to word information. In particular we attempted to have good coverage for features and feature extraction regions that expected to play a role in the three focussed analyses mentioned in the Introduction: detection of questions, agreements, and incomplete utterances. Based on the literature on question intonation (Vaissière, 1983; Haan et al., 1997a; van Heuven et al., 1997), we expected questions to show rising F0 at the end of the utterance, particularly for declarative and yes-no questions. Thus, F0 should be a helpful cue for distinguishing questions from other long DAs such as statements. Many incomplete utterances give the impression of being cut off prematurely, so the prosodic behaviour at the end of such an utterance may be

similar to that of the middle of a normal utterance. Specifically, energy can be expected to be higher at the end of an abandoned utterance compared to a completed one. In addition, unlike most completed utterances, the F0 contour at the end of the utterance is neither rising nor falling. For these reasons RMS energy and F0 were calculated additionally within regions near the end of the utterance. We expected backchannels to differ from agreements by the amount of effort used in speaking. Backchannels function to acknowledge another speaker's contributions without taking the floor, whereas agreements assert an opinion. We therefore expected agreements to have higher energy, greater F0 movement, and a higher likelihood of accents and boundary tones.

Duration and pause features. Duration was expected to be a good cue for discriminating statements and questions from DAs functioning to manage the dialog (e.g. backchannels), although this difference is also encoded to some extent in the language model. In addition to the duration of the utterance in seconds, we included features correlated with utterance duration but based on frame counts conditioned on the value of other feature types, as shown in Table 2.

Table 2: Duration Features

Feature Name	Description
Duration	
ling_dur	duration of utterance (linguistically-segmented)
Duration-pause	
ling_dur_minus_min10pause	ling_dur minus sum of duration of all pauses of at least 100 msec
cont_speech_frames	number of frames in continuous speech regions ($>$ 1 sec, ignoring pauses $<$ 100 msec)
Duration-correlated F0-based counts	
f0_num_utt	number of frames with F0 values in utterance (prob_voicing=1)
f0_num_good_utt	number of F0 values above f0_min (f0_min = .75*f0_mode)
regr_dur	duration of F0 regression line (from start to end point, includes voiceless frames)
regr_num_frames	number of points used in fitting F0 regression line (excludes voiceless frames)
numacc_utt	number of accents in utterance from event recognizer
numbound_utt	number of boundaries in utterance from event recognizer

The duration-pause set of features computes duration ignoring pause regions. Such features may be

useful if pauses are unrelated to DA-classification. (If pauses are relevant however, this should be captured by the pause features described in the next section). The F0-based count features reflect either the number of frames or recognized intonational events (accents or boundaries) based on F0 information (see F0 features, below). The first four of these features capture time in speaking using knowledge about the presence and location of voiced frames, which may be more robust for our data than relying on pause locations from the alignments. The last two features are intended to capture the amount of information in the utterance, by counting accents and phrase boundaries. Duration-normalized versions of many of these features are included under their respective feature type in the following sections.

Pause features. To address the possibility that hesitation could provide a cue to the type of DA, we included features intended to reflect the degree of pausing, as shown in Table 3. To obtain pause locations we used information available from forced-alignments; however this was only for convenience (the alignment information was in our database for other purposes). In principle, pause locations can be detected by current recognizers with high accuracy without knowledge of the words. Pauses with durations below 100 milliseconds (10 frames) were excluded since they are more likely to reflect segmental information than hesitation. Features were normalized to remove the inherent correlation with utterance duration. The last feature was included to provide a more global measure of pause behavior, including pauses during which the other speaker was talking. The measure counts only those speech frames occurring in regions of at least one second of continuous speaking. The window was run over the conversation side, writing out a binary value for each frame; the feature was then computed based on the frames within a particular DA.

Table 3: Pause Features

Feature Name	Description
min10pause_count_n_ldur	number of pauses of at least 10 frames (100 msec) in utterance, normalized by duration of utterance
total_min10pause_dur_n_ldur	sum of duration of all pauses of at least 10 frames in utterance, normalized by duration of utterance
mean_min10pause_dur_utt	mean pause duration for pauses of at least 10 frames in utterance
mean_min10pause_dur_ncv	mean pause duration for pauses of at least 10 frames in utterance, normalized by same in convside
cont_speech_frames_n	number of frames in continuous speech regions ($>$ 1 sec, ignoring pauses $<$ 10 frames) normalized by duration of utterance

F0 features. F0 features, shown in Table 4, included both raw and regression values based on frame-level F0 values from ESPS/Waves+.

Table 4: F0 Features

Feature Name	Description
f0_mean_good_utt	mean of F0 values included in f0_num_good_utt
f0_mean_n	difference between mean F0 of utterance and mean F0 of convside for F0 values > f0_min
f0_mean_ratio	ratio of F0 mean in utterance to F0 mean in convside
f0_mean_zcv	mean of good F0 values in utterance normalized by mean and st dev of good F0 values in convside
f0_sd_good_utt	st dev of F0 values included in f0_num_good_utt
f0_sd_n	log ratio of st dev of F0 values in utterance and in convside
f0_max_n	log ratio of max F0 values in utterance and in convside
f0_max_utt	maximum F0 value in utterance (no smoothing)
max_f0_smooth	maximum F0 in utterance after median smoothing of F0 contour
f0_min_utt	minimum F0 value in utterance (no smoothing); can be below f0_min
f0_percent_good_utt	ratio of number of good F0 values to number of F0 values in utterance
utt_grad	least-squares all-points regression over utterance
pen_grad	least-squares all-points regression over penultimate region
end_grad	least-squares all-points regression over end region
end_f0_mean	mean F0 in end region
pen_f0_mean	mean F0 in penultimate region
abs_f0_diff	difference between mean F0 of end and penultimate regions
rel_f0_diff	ratio of F0 of end and penultimate regions
norm_end_f0_mean	mean F0 in end region normalized by mean and st dev of F0 from convside
norm_pen_f0_mean	mean F0 in penultimate region normalized by mean and st dev from convside
norm_f0_diff	difference between mean F0 of end and penultimate regions, normalized by mean and st dev of F0 from convside
regr_start_f0	first F0 value of contour, determined by regression line analysis
finalb_amp	amplitude of final boundary (if present), from event recognizer
finalb_label	label of final boundary (if present), from event recognizer
finalb_tilt	tilt of final boundary (if present), from event recognizer
numacc_n_ldur	number of accents in utterance from event recognizer, normalized by duration of utterance
numacc_n_rdur	number of accents in utterance from event recognizer, normalized by duration of F0 regression line
numbound_n_ldur	number of boundaries in utterance from event recognizer, normalized by duration of utterance
numbound_n_rdur	number of boundaries in utterance from event recognizer, normalized by duration of F0 regression line

To capture overall pitch range, mean F0 values were calculated over all voiced frames in an utterance. To normalize differences in F0 range over speakers, particularly across genders, utterance-level values were normalized with respect to the mean and standard deviation for F0 values measured over the whole

conversation side. F0 difference values were normalized on a log scale. The standard deviation in F0 over an utterance was computed as a possible measure of expressiveness over the utterance. Minimum and maximum F0 values, calculated after median smoothing to eliminate spurious values, were also included for this purpose.

We also included parallel measures that used only "good" F0 values, or values above a threshold ("f0_min") estimated as the bottom of a speaker's natural F0 range. The f0_min can be calculated in two ways. For both methods, a smoothed histogram of all the calculated F0 values for a conversation side is used to find the F0 mode. The true f0_min comes from the minimum F0 value to the left of this mode. Because the histogram can be flat or not sufficiently smoothed, our algorithm may be fooled into choosing a value greater than the true minimum. A simpler way to estimate f0_min takes advantage of the fact that values below the minimum typically result from pitch halving. Thus, a good estimate of f0_min is to take the point at 0.75 times the F0 value at the mode of the histogram. This measure closely approximates the true f0_min, and is more robust for use with the Switchboard data.[1] The percentage of "good" F0 values was also included to measure (inversely) the degree of creaky-voice or vocal fry.

The rising/falling behavior of contours is a good cue to their utterance type. We investigated a number of ways to measure this behaviour. To measure overall slope, we calculated the gradient of a least-squares fit regression line for the F0 contour. While this gives an adequate measure for the overall gradient of the utterance, it is not always a good indicator of the type of rising/falling behavior we are most interested in. Rises at the end can be swamped by the declination of the part of the contour preceding this, and hence the overall gradient for a contour can be falling. We therefore marked two special regions at the end of the contour, corresponding to the last 200ms ("end" region) and the previous 200ms to that ("penultimate" region). For each of these regions we measured the mean F0 and gradient, and used the differences between these as features. The starting value in the regression line was also included as a potential cue to F0 register (the actual first value is prone to F0 measurement error).

In addition to these F0 features, we also included intonational-event features, or features intended to

[1] We thank David Talkin for suggesting this method.

capture local pitch accents and phrase boundaries. The event features obtained using the event recognizer described in Taylor et al. (1997b). The event detector uses a HMM approach to provide an intonational segmentation of an utterance, which gives the locations of pitch accents and boundary tones. When compared to human intonation transcriptions of Switchboard, this system correctly identifies 64.9% of events, but has a high false alarm rate, resulting in an accuracy of 31.7%.

Energy features. We included two types of energy features, as shown in Table 5. The first set of features was computed based on standard root mean square (rms) energy. Because our data were recorded from telephone handsets with various noise sources (background noise as well as channel noise) we also included a signal-to-noise ratio (SNR) feature to try to capture the energy from the speaker. SNR values were calculated using the SRI recognizer with a Switchboard-adapted front-end (Neumeyer and Weintraub, 1994; Neumeyer and Weintraub, 1995). Values were calculated over the entire conversation side, and those extracted from regions of speech were used to find a cumulative distribution function (CDF) for the conversation. The frame-level SNR values were then represented by their CDF value in order to normalize the SNR values across speakers and conversations.

Table 5: Energy Features

Feature Name	Description
utt_nrg_mean	mean RMS energy in utterance
abs_nrg_diff	difference between mean RMS energy of end and penultimate regions
end_nrg_mean	mean RMS energy in end region
norm_nrg_diff	normalized difference between mean RMS energy of end and penultimate regions
rel_nrg_diff	ratio of mean RMS energy of end and penultimate regions
snr_mean_utt	mean signal-to-noise ratio (CDF value) in utterance
snr_sd_utt	st dev of signal-to-noise ratio values (CDF values) in utterance
snr_diff_utt	difference between maximum SNR and minimum SNR in utterance
snr_min_utt	st dev of signal-to-noise ratio values (CDF values) in utterance
snr_max_utt	maximum signal-to-noise ratio values (CDF values) in utterance

Speaking rate ("enrate") features. We were also interested in overall speaking rate. However we

needed a measure that could be run directly on the signal. For this purpose, we experimented with a signal processing measure, "enrate," which is currently under development by Morgan et al. (1997). This measure runs directly on the speech signal, and has been shown to correlate moderately with lexical measures of speaking rate. The measure can be run over the entire signal, but because it uses a large window, values are less meaningful if significant pause time is included in the window. Since our speakers were recorded continuously, we had long pause regions in our data (mainly times during which the other speaker was talking). Based on advice from the developers, we applied the measure to certain stretches of speech of minimum duration without excessive pauses.

We calculated frame-level values over a two second speech interval. The enrate value was calculated for a 25ms frame window with a window step hop of 0.1 second. Output values were calculated for 10ms frames to correspond to other measurements. We included pauses of less than 1 second and ignored speech regions of less than one second, where pause locations were determined as described earlier.

If the end of a speech segment was approaching, meaning that the 2 second window could not be filled, no values were written out. The enrate values corresponding to particular utterances were then extracted from the conversation side values. This way, if utterances were adjacent, information from surrounding speech regions could be used to get enrate values for the beginnings and ends of utterances which normally would not fill the 2 second speech window. Features computed for use in tree-building are listed in Table 6.

Table 6: Speaking Rate Features

Feature Name	Description
mean_enr_utt	mean of enrate values in utterance
mean_enr_utt_norm	mean_enr_utt normalized by mean enrate in conversation-side
stdev_enr_utt	st dev of enrate values in utterance
min_enr_utt	minimum enrate value in utterance
max_enr_utt	maximum enrate value in utterance

Gender features. We also included gender features. This was not a main focus of our study, but as a way to check the effectiveness of our F0 normalizations we included the gender of the speaker. It is

also possible, however, that features could be used differently by men and women, even after appropriate normalization for pitch range differences. We also included the gender of the listener to check for a possible sociolinguistic interaction between the ways in which speakers employ different prosodic features and the conversational dyad.

Decision Tree Classifiers

For our prosodic classifiers, we used CART-style decision trees (Breiman et al., 1983). Decision trees allow combination of discrete and continuous features, and can be inspected to gain an understanding of the role of different features and feature combinations.

We downsampled our data to obtain an equal number of datapoints in each class. Although a drawback to downsampling is a loss of power in the analyses due to fewer datapoints, downsampling was warranted for two important reasons. First, as noted earlier, the distribution of frequencies for our DA classes was severely skewed. Because decision trees split according to an entropy criterion, large differences in class sizes wash out any effect of the features themselves, causing the tree not to split. By downsampling to equal class priors we assure maximum sensitivity to the features. A second motivation for downsampling was that by training our classifiers on a uniform distribution of DAs, we facilitated integration with other knowledge sources (see section on Integration).

After finishing expanding the tree with questions, the tree-growing algorithm used a ten-fold cross-validation procedure to avoid overfitting the training data. Leaf nodes were successively pruned if they failed to reduce the entropy in the cross-validation procedure.

We report tree performance using two metrics, accuracy and efficiency. Accuracy is the number of correct classifications divided by the total number of samples. Accuracy is based on hard decisions; the classification is that class with the highest probability. Because we downsample to equal class priors, the chance performance for any tree with N classes is 100/N%. For any particular accuracy level, there is a trade-off between recall and false alarms. In the real world there may well be different costs to a false positive versus a false negative in detecting a particular utterance type. In the absence of any model of how such costs would be assigned for our data, we report results assuming equal costs to these errors for our

downsampled trees.

Efficiency measures the relative reduction in entropy between the prior class distribution and the posterior distribution predicted by the tree. Two trees may have the same classification accuracy, but the tree which more closely approximates the probability distributions of the data (even if there is no effect on decisions) has higher efficiency (lower entropy). Although accuracy and efficiency are typically correlated, the relationship between the measures is not strictly monotonic since efficiency looks at probability distributions and accuracy looks only at decisions.

DA Classification from Word Sequences

Two methods were used for classification of DAs from word information. For experiments using the correct words W, we needed to compute the likelihoods $P(W|U)$ for each DA or utterance type U, i.e., the probability with which U generates the word sequence W. The predicted DA type would then be the one with maximum likelihood. To estimate these probabilities, we grouped the transcripts of the training corpus by DA, and trained a standard trigram language model using backoff smoothing (Katz, 1987) for each DA. This was for done for the original 42 DA categories, yielding 42 DA-specific language models. Next, for experiments involving a DA class C comprising several of the original DAs U_1, U_2, ..., U_n, we combined the DA likelihoods in a weighted manner:

$$P(W|C) = P(W|U_1)P(U_1|C) + \ldots + P(W|U_n)P(U_n|C)$$

Here, $P(U_1|C)$, ..., $P(U_n|C)$ are the relative frequencies of the various DAs within class C.

For experiments involving (necessarily imperfect) automatic word recognition, we were given only the acoustic information A. We therefore needed to compute acoustic likelihoods $P(A|U)$, i.e., the probability that utterance type U generates the acoustic manifestation A. In principle, this can be acomplished by considering all possible word sequences W that might have generated the acoustics A, and summing over them:

$$P(A|U) = \sum_W P(A|W)P(W|U)$$

Here $P(W|U)$ is estimated by the same DA-specific language models as before, and $P(A|W)$ is the acoustic score of a speech recognizer, expressing how well the acoustic observations match the word sequence W. In practice, however, we could only consider a finite number of potential word hypotheses W; in our experiments we generated the 2500 most likely word sequences for each utterance, and carried out the above summation over only those sequences. The recognizer used was a state-of-the-art HTK large vocabulary recognizer, which nevertheless had a word error rate of 41% on the test corpus.[2]

Integration of Knowledge Sources

To use multiple knowledge sources for DA classification, i.e., prosodic information as well as other, word-based evidence, we combined tree probabilities $P(U|F)$ and word-based likelihoods $P(W|U)$ multiplicatively. This approach can be justified as follows. The likelihood-based classifier approach dictates choosing the DA with the highest likelihood based on both the prosodic features F and the words W, $P(F, W|U)$. To make the computation tractable, we assumed that the prosodic features are independent of the words once conditioned on the DA. We recognize however that this assumption is a simplification. Our prosodic model averages over all examples of a particular DA; it is "blind" to any differences in prosodic features that correlate with word information. For example, statements about a favorite sports team use different words than statements about personal finanace, and the two different types of statements tend to differ prosodically (for example in animation level as reflected by overall pitch range). In future work such differences could potentially be captured using more sophisticated models designed to capture semantic or topic information. For practical reasons, however, we consider our prosodic models independent of the words once conditioned on the DA, i.e.:

$$\begin{aligned} P(F, W|U) &= P(W|U)P(F|W, U) \\ &\approx P(W|U)P(F|U) \end{aligned}$$

[2]Note that the summation over multiple word hypotheses is preferable to the more straighforward approach of looking at only the one best hypothesis and treating it as the actual words for the purpose of DA classification.

$$\propto \quad P(W|U)P(U|F)$$

The last line is justified because as noted earlier, we trained the prosodic trees on downsampled data or a uniform distribution of DA classes. Therefore the posterior probabilities, $P(U|F)$, it produces are proportional to the required likelihoods $P(F|U)$. Overall, this justifies multiplying $P(W|U)$ and $P(U|F)$.[3]

Training and Test sets

We partitioned the available data into three subsets for training and test purposes. The three subsets were not only disjoint but also shared no speakers. The *training set* (TRN) contained 1794 conversation sides; its acoustic waveforms were used to train decision trees, while the corresponding transcripts served as training data for the statistical language models used in word-based DA classification. The *held-out set* (HLD) contained 436 conversation sides; it was used to test tree performance, as well as true-word based DA classification. A much smaller *development test set* (DEV) consisting of 38 matched conversation sides (19 conversations) was used to perform experiments involving automatic word-recognition, as well as the corresponding prosody and true-word based experiments.[4] The TRN and HLD sets contained single, unmatched conversation sides, but since no discourse context was required for the studies reported here this was not a problem. The three corpus subsets with their statistics are summarized in Table 7.

Table 7: Summary of corpus subsets for training and testing

Name	Description	Sides	Utterances	Words
TRN	Training set	1794	166K	1.2M
HLD	Held-out test set	436	32K	231K
DEV	Development test set	19	4K	29K

[3] In practice we needed to adjust the dynamic ranges of the two probability estimates by finding a suitable exponential weight λ, to make

$$P(F, W|U) \propto P(W|U)P(F|U)^{\lambda}$$

The weight λ was found by optimizing on held-out data.

[4] The DEV set was so called because of its role in the WS97 projects which focussed on word recognition

RESULTS AND DISCUSSION

We first examine results of the prosodic model for a seven-way classification involving all DAs. We then look to results from a words-only analysis, to discover potential subtasks for which prosody could be particularly helpful. The analysis reveals that even if correct words are available, certain DAs tend to be misclassified. We examine the potential role of prosody for three such subtasks: (1) the detection of questions; (2) the detection of agreements; and (3) the detection of incomplete utterances. In all analyses we seek to understand the relative importance of different features and feature types, as well as to detemine whether integrating prosodic information with a language model can improve classification performance overall.

Seven-Way Classification

We applied the prosodic model first to a seven-way classification task for the full set of DAs: Statements, Questions, Incomplete utterances, Backchannels, Agreements, Appreciations, and Other. Note that "Other" is a catch-all class representing a large number of heterogenous DAs that occurred infrequently in our data. Therefore we do not expect this class to have consistent features, but rather to be distinguished to some extent based on feature consistencies within the other six classes. As described in the Method section, data were downsampled to equal class sizes to avoid confounding results information from prior frequencies of each class. Because there are seven classes, chance accuracy for this task is 100/7% or 14.3%. For simplicity, we assumed equal cost to all decision errors, i.e. to all possible confusions among the seven classes.

A tree built using the full database of features described earlier allows a classification accuracy of 41.15%. This gain in accuracy is highly significant by a binomial test; $p < .0001$. If we are interested in probability distributions rather than decisions, we can look at the efficiency of the tree, or the relative reduction in entropy over the prior distribution. By using the all-features prosodic tree for this seven-way classification, we reduce the number of bits necessary to describe the class of each datapoint by 16.8%.

The all-features tree is large (52 leaves), making it difficult to interpret the tree directly. In such cases we find it useful to summarize the overall contribution of different features using a measure 'feature usage,"

which is proportional to the number of times a feature was queried in classifying the datapoints. The measure thus accounts for the position of the feature in the tree: features used higher in the tree have greater usage values than those lower in the tree since there are more datapoints at the higher nodes. The measure is normalized to sum to 1.0 for each tree. Table 8 lists usage by feature type.

Table 8: Feature Usage for Main Feature Types in Seven-Way Classification

Feature Type	Usage (%)
Dur	0.554
F0	0.126
Pau	0.121
Nrg	0.104
Enr	0.094

Table 8 indicates that when all features are available, a duration-related feature is used in more than half of the queries. And notably, gender features are not used at all; this supports the earlier hypothesis that, as long as features are appropriately normalized, it is reasonable to create gender-independent prosodic models for these data. Individual feature usage, as shown in Table 9, reveals that the raw duration feature (ling_dur)—which is a "free" measure in our work since we assumed locations of utterance boundaries—accounted for only 14% of the queries in the tree; the remaining portion of the 55% accounted for by duration features were those associated with the computation of F0- and pause-related information. Thus the power of duration for the seven-way classification comes largely from measures involving computation of other prosodic features. The most-queried feature, regr_num_frames (the number of frames used in computing the F0 regression line) may be better than other duration measures at capturing actual speech portions (as opposed to silence or nonspeech sounds), and may be better than other F0-constrained duration measures (e.g. f0_num_good_utt) due to a more robust smoothing algorithm. We can also note that the high overall rate of F0 feature given in Table 9 represents a summation over many different individual features.

Table 9: Feature Usage for Seven-Way (All DAs) Classification

Feature Type	Feature	Usage (%)
Dur	regr_num_frames	0.180
Dur	ling_dur	0.141
Pau	cont_speech_frames_utt_n	0.121
Enr	stdev_enr_utt	0.081
Enr	ling_dur_minus_min10pause	0.077
Enr	cont_speech_frames_utt	0.073
Nrg	snr_max_utt	0.049
Nrg	snr_mean_utt	0.043
Dur	regr_dur	0.041
F0	f0_mean_zcv	0.036
F0	f0_mean_n	0.027
Dur	f0_num_good_utt	0.0214
Dur	f0_num_utt	0.019
F0	norm_end_f0_mean	0.017
F0	numacc_n_rdur	0.016
F0	f0_sd_good_utt	0.015
Enr	mean_enr_utt	0.009
F0	f0_max_n	0.006
Nrg	snr_sd_utt	0.006
Nrg	rel_nrg_diff	0.005
Enr	mean_enr_utt_norm	0.004
F0	regr_start_f0	0.003
F0	finalb_amp	0.003

Since we were also interested in feature importance, a number of individual trees were built using the leave-one-out method, in which the feature list is systematically modified and a new tree is built for each subset of allowable features. It was not feasible to leave out individual features because of the large set of features used; we therefore left out groups of features corresponding to the feature types as defined in the Method section. We also included a matched set of "leave-one-in" trees for each of the feature types (i.e. trees for which all *other* feature types were removed), and a single leave-two-in tree, built *post hoc* which made available the two feature types with highest accuracy from the leave-one-in analyses. Note that the defined feature lists specify the features *available* for use in building a particular prosodic model; whether or not features are *actually* used requires inspection of the resulting tree. Figure 1 shows results for the

set of leave-one-out and leave-one-in trees, with the all-features tree provided for comparison purposes. The upper graph indicates accuracy values; the lower graph shows efficiency values. Each bar indicates a separate tree.

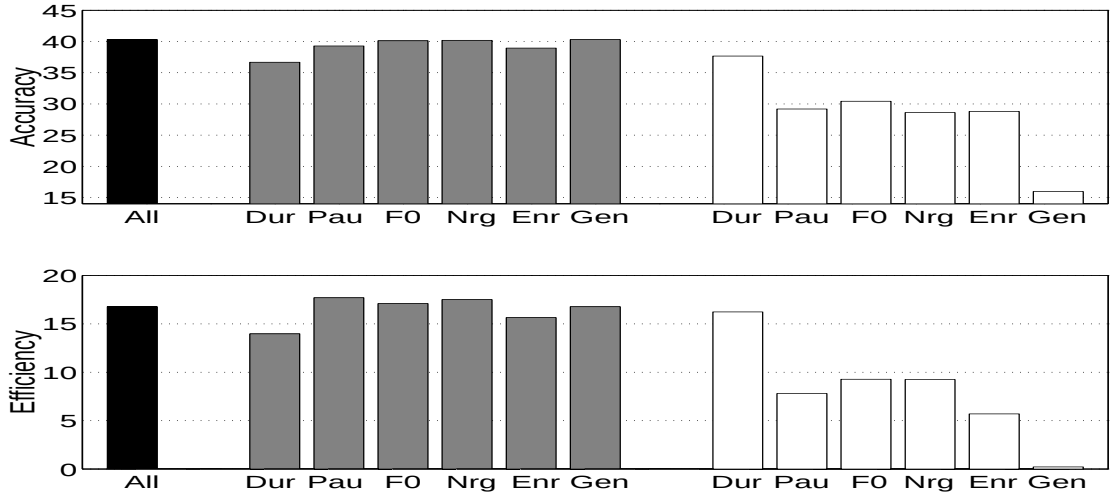

Figure 1: Performance of prosodic trees using different feature sets for the classification of all seven DAs (Statements, Questions, Incomplete Utterances, Backchannels, Agreements, Appreciations, Other). N for each class=391. Chance accuracy = 14.3%. Gray bars=exclude feature type; white bars=include only feature type. Dur=Duration, Pau=Pause, F0=Fundamental frequency, Nrg=Energy, Enr=Enrate (speaking rate), Gen=Gender features.

We first tested whether there was any significant loss in leaving out a feature type, by doing pairwise comparisons between the all-features tree and each of the leave-one-out trees.[5] Although trees with more features to choose from typically perform better than those with fewer features, additional features can hurt performance. The greedy algorithm used cannot look ahead to determine the optimal overall model, but rather seeks to maximize entropy reduction locally at each split. This limitation of decision trees is another motivation for conducting the leave-one-out analyses. Since we cannot predict the direction of change for different feature sets, comparison on tree results are conducted using two-tailed tests.

[5]To test whether one tree (A) was significantly better than another (B), we counted the number of test instances on which A and B differed, and on how many A was correct but B was not; we then applied a Sign test to these counts.

Results showed that the only two feature types whose removal caused a significant reduction in accuracy were duration ($p < 0.0001$) and enrate ($p < 0.05$). The enrate-only tree however yields accuracies on par with other feature types whose removal did not affect overall performance; this suggests that the contribution of enrate in the overall tree may be through interactions with other features. All of the leave-one-in trees were significantly less accurate than the all-features tree; although the tree using only duration achieved an accuracy close to that of the all-features tree, it was still significantly less accurate by a Sign test ($p < 0.01$). Adding F0 features (the next best feature set in the leave-one-in trees) did not significantly improve accuracy over the duration-only tree alone, suggesting that for this task the two feature types are highly correlated. Nevertheless, each of the leave-one-in trees, all feature types except gender yielded accuracies significantly above chance by a binomial test ($p < .0001$ for the first five trees). The gender-only tree was slightly better than chance by either a one- or a two-tailed test,[6] however this was most likely due to a difference in gender representation across classes.

Taken together, these results suggest that there is considerable redundancy in the features for DA classification, since removing one feature type at a time (other than duration) makes little difference to accuracy. Results also suggest however that features are not perfectly correlated; there must be considerable interaction among features in classifying DAs, because trees using only individual feature types are significantly less accurate than the all-features tree.

Finally, duration is clearly of primary importance to this classification. This is not surprising, as the task involves a seven-way classification including longer utterances (such as statements) and very brief ones (such as backchannels like "uh-huh"). Two questions of further interest regarding duration, however, are: (1) will a prosody model that uses mostly duration add anything to a language model (in which duration is implicitly encoded); and (2) is duration useful for other tasks involving classification of DAs similar in length. Both questions are addressed in the following sections.

As just discussed, the all-features tree (as well as others including only subsets of feature types) provide significant information for the seven-way classification task. Thus if one were only to use prosodic

[6] It is not clear here whether a one- or two-tailed test is more appropriate. Trees typically should not do worse than chance; however because they minimize entropy, and not accuracy, the accuracy can fall slightly below chance.

information (no words or context), this is the level of performance resulting for the case of equal class frequencies. To explore whether the prosodic information could be of use when lexical information is also available, we integrated the tree probabilities with likelihoods from our DA-specific trigram language models built from the same data. For simplicity, integration results are reported only for the all-features tree in this and all further analyses, although as noted earlier this is not guaranteed to be the optimal tree.

Since our trees were trained with uniform class priors, we combined tree probabilities $P(U|F)$ with the word-based likelihoods $P(W|U)$ multiplicatively, as described in the Integration section, using a weighting factor found by optimizing on held out data. The integration was performed separately for each of our two test sets (HLD and DEV), and within the DEV set for both transcribed and recognized words. Results are shown in Table 10. Classification performance is shown for each of the individual classifiers, as well as for the optimized combined classifier.

Table 10: Accuracy of Individual and Combined Models for Seven-Way Classification

Knowledge Source	HLD Set true words	DEV Set true words	DEV Set N-best output
samples	2737	287	287
chance (%)	14.29	14.29	14.29
tree (%)	41.15	38.03	38.03
words (%)	67.61	70.30	58.77
words+tree (%)	69.98	71.14	60.12

As shown, for all three analyses, adding information from the tree to the words model improved classification accuracy. Although the gain appears modest in absolute terms, for the HLD test set it was highly significant by a Sign test,[7] $p < .001$. For the smaller DEV test set, the improvements did not reach significance; however the pattern of results suggests that this is likely to be due to the small sample size. It is also the case that the tree model does not perform as well for the DEV as the HLD set; this is not attributable to small sample size, but rather to a mismatch between the DEV set and the training data involving how data were segmented, as noted in the Method section. The mismatch in particular affects duration features,

[7] One-tailed, because model integration assures no loss in accuracy.

which were important in these analyses as discussed earlier. Nevertheless, word-model results are lower for N-best than for true words in the DEV data while by definition the tree results stay the same. We see that accordingly, integration provides a larger win for the recognized than the true words. Thus we would expect results for recognized words for the HLD set (if they could be obtained) should show an even larger win than the win observed for the true words in that set.

These results provide an answer to one of the questions posed in the previous section: does prosody provide an advantage over words if the prosody model uses mainly duration? The results indicate that the answer is yes. Although the number of words in an utterance is highly correlated with duration, and word counts are represented implicitly by the probability of the end-of-utterance marker in a language model, a duration-based tree model still provides added benefit over words alone. One reason may be that duration (reflected by the various features we included) is simply a better predictor of DA than is word count. Another independent possibility is that the advantage from the tree model comes from its ability to directly and iteratively threshold feature values.

DA Confusions Based on Word Information

Next we explored additional tasks for which prosody could aid DA classification. Since our trees allow N-ary classification, the logical search space of possible tasks was too large to explore systematically. We therefore looked to the language model to guide us in identifying particular tasks of interest. Specifically, we were interested in DAs that tended to be misclassified even given knowledge of the true words. We therefore examined the pattern of confusions made when our seven DAs were classified using the language model alone. Results are shown in Figure 2. Each subplot represents the data for one actual DA.[8] Bars reflect the normalized rate at which the actual DA was classified as each of the seven possible DAs, in each of the the three test conditions (HLD, DEV-true, and DEV-Nbest).

[8] Due to the heterogeneous makeup of the "other" DA class *per se*, we were not particularly interested in its pattern of confusions and hence the graph for that data is not shown.

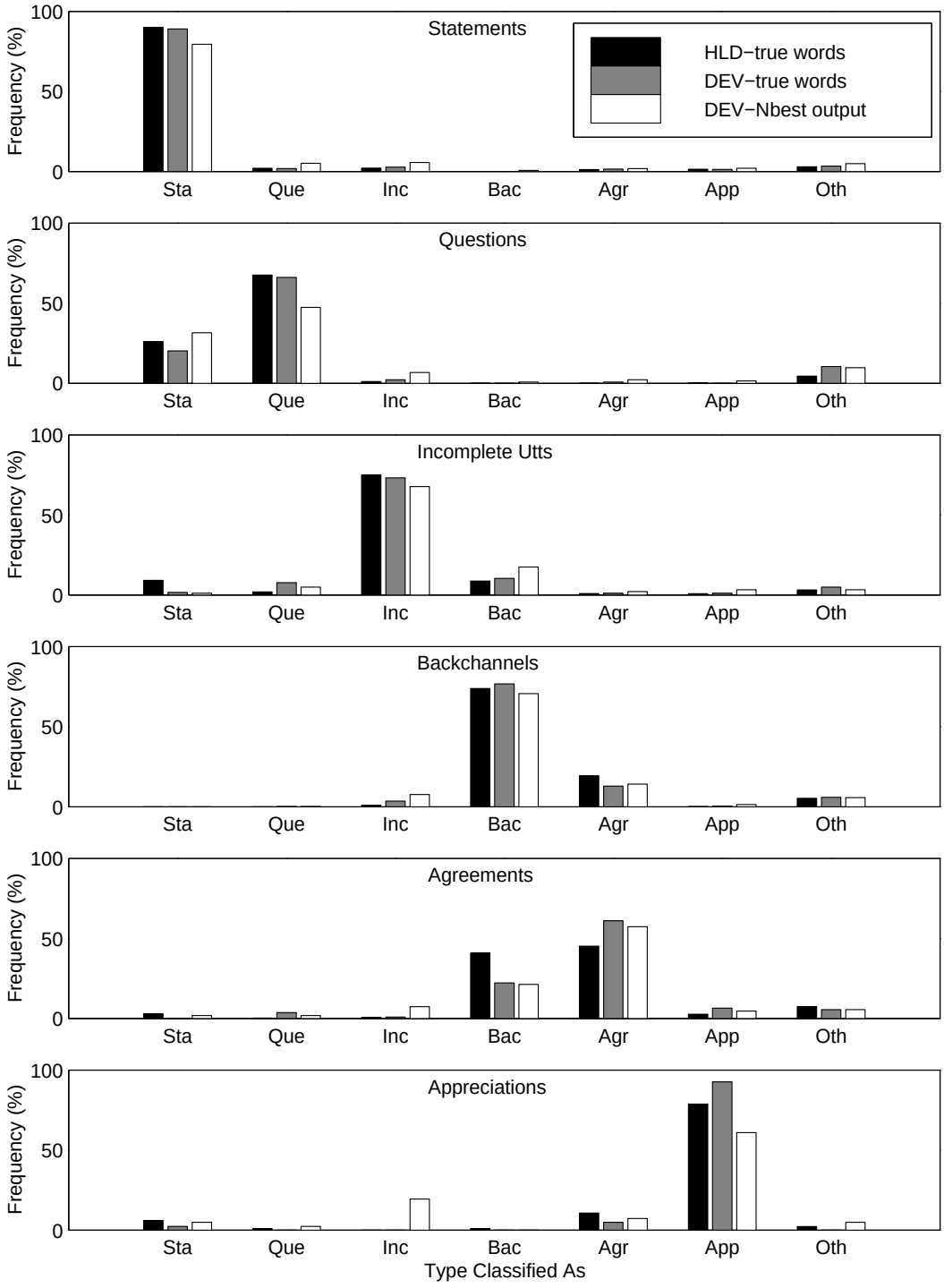

Figure 2: Classification of DAs based on word trigrams only, using three different test sets.

As shown, classification is excellent for the statement class, with few misclassifications even when only the recognized words are used.[9] For the remaining DAs however, misclassifications occur at considerable rates. Classification of questions is a case in point: even using true words, questions are often misclassified as statements (but not vice versa), and this pattern is exaggerated when testing on recognized as opposed to true words. The asymmetry is partially attributable to the presence of declarative questions. The effect associated with recognized words appears to reflect a high rate of missed initial "do" in our recognition output, as discovered in independent error analyses (Jurafsky et al., 1997a). For both statements and questions however, there is little misclassification involving the remaining classes. This probably reflects the length distinction as well as the fact that most of the propositional content in our corpus occurred in statements and questions, whereas other DAs generally served to manage the communication—a distinction likely to be reflected in the words. Thus, our first subtask will be to examine the role of prosody in the classification of statements and questions.

A second problem visible in Figure 2 is the detection of incomplete utterances. Even using true words, classification of these DAs is at only 75.0% accuracy. Knowing whether or not a DA is complete would be particularly useful for both language modeling and understanding. Since the misclassifications are distributed over the set of DAs, and since logically any DA can have an incomplete counterpart, our second subtask will be to classify a DA as either incomplete or not-incomplete (all other DAs).

A third notable pattern of confusions involves backchannels and explicit agreements. This is not surprising, since the two classes share words such as "yeah" and "right". In this case, the confusions are considerable in both directions, but more marked for the case of agreements. As mentioned in the Method section, some of these cases may involve utterances that were mislabeled because labelers used only the transcripts. However, for any mislabeled cases we would expect no improvement by adding prosody, since we would also need to match the (incorrect) transcriber labels. Thus any gain from prosody would be likely to reflect a contribution for correctly labeled cases; we will therefore examine backchannels and agreements

[9]The high classification rate for statements by word information was a prime motivation for downsampling our data in order to examine the inherent contribution of prosody, since as noted in the Method section, statements make up most of the data in this corpus.

as our third classification subtask.

Subtask 1: Detection of questions

As just illustrated in the previous section, questions in our corpus were often misclassified as statements based on words alone. Based on the literature, we hypothesized that prosodic features, particularly those capturing the final F0 rise typical of some question types in English, could play a role in reducing the rate of misclassifications.

To investigate the hypothesis, we built a series of classifiers using only question and statement data. We first examined results for an all-features tree, shown in Figure 3. The tree yields an accuracy of 74.21%, which is significantly above the chance level of 50% by a binomial test, $p < .0001$; it reduces the number of bits necessary to describe the class of each datapoint by 20.9%.

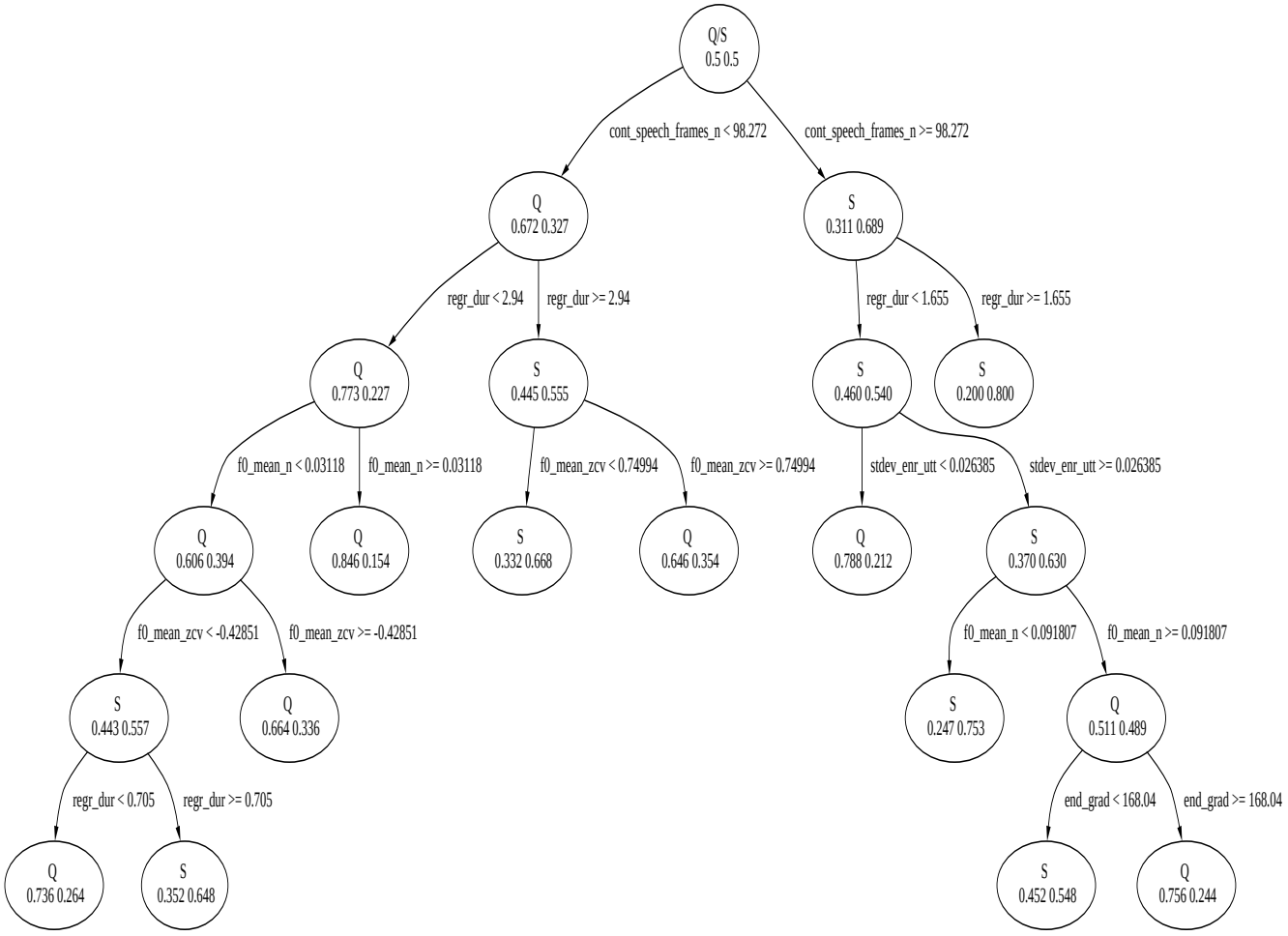

Figure 3: Decision tree for the classification of Statements (S) and Questions (Q).

Feature importance. The feature usage of the tree is summarized in Table 11. As predicted, F0 features help differentiate questions from statements, and in the expected direction (questions have higher F0 means, and higher end gradients than statements). What was not obvious at the outset is the extent to which other features cue this distinction. In the all-features tree, F0 features comprise only about 28% of the total queries. Two other features, regr_dur, and cont_speech_frames, are each queried more often than the F0 features together. Questions are shorter in duration (from starting to ending voiced frame) than statements. They also have a lower percentage of frames in continuous speech regions than statements.

Further inspection suggests that the role of the pause feature in this case (and also most likely for the seven-way classification discussed earlier) is in the form of an "external" prosodic feature. As noted in the Method section, for this feature the minimum threshold for a continuous speech region is one second. Since the feature was run over the conversation side, and since nonspeech regions for one speaker are correlated with speech regions by the other, the feature indirectly captures some turn change information. Specifically, if a DA is followed directly by more speech from the same speaker, the one-second window continues across the end of the first DA. In this case all frames in the window count toward the value of the feature. If however the DA is *not* directly followed by more speech from the same speaker, the last full window will end before the final frame of the DA, and fewer frames will count toward the value of the feature. The same explanation applies to DA onsets. In our data, questions and statements were about equally likely to have a turn boundary on only one side (in both cases with turn boundaries more likely following than preceding the DA); however questions were more than three times as likely as statements to have a turn boundary on both sides. This difference is likely to be captured to some extent by the pause feature.

Table 11: Feature Usage for Classification of Questions and Statements

Feature Type	Feature	Usage (%)
Dur	regr_dur	0.332
Pau	cont_speech_frames_n	0.323
F0	f0_mean_n	0.168
F0	f0_mean_zcv	0.088
Enr	stdev_enr_utt	0.065
F0	end_grad	0.024

To further examine the role of features we built additional trees using partial feature sets. Results are summarized in Figure 4. As suggested by the leave-one-out trees, there is no significant effect on accuracy when any one the feature types is removed. Although we predicted that questions should differ from statements mainly by intonation, results indicate that a tree with no F0 features achieves the same accuracy as a tree with all features for the present task. Removal of all pause features, which resulted in the largest

drop in accuracy, yields a tree with an accuracy of 45.53%, which is not significantly different from that of the all-features tree ($p = .2111$, n.s.). Thus if any feature type is removed, other feature types compensate to provide the same overall accuracy. However, it is not the case that the main features used are perfectly correlated, with one substituting in when another has been removed. Inspection of the leave-one-out tree reveals that upon removal of a feature type, new features (features, and feature types, that never appeared in the all-features tree) are used. Thus: (1) there is a high degree of redundancy in the features that differentiate questions and statements; and (2) the relationship among these features and the allowable feature sets for tree building is complex.

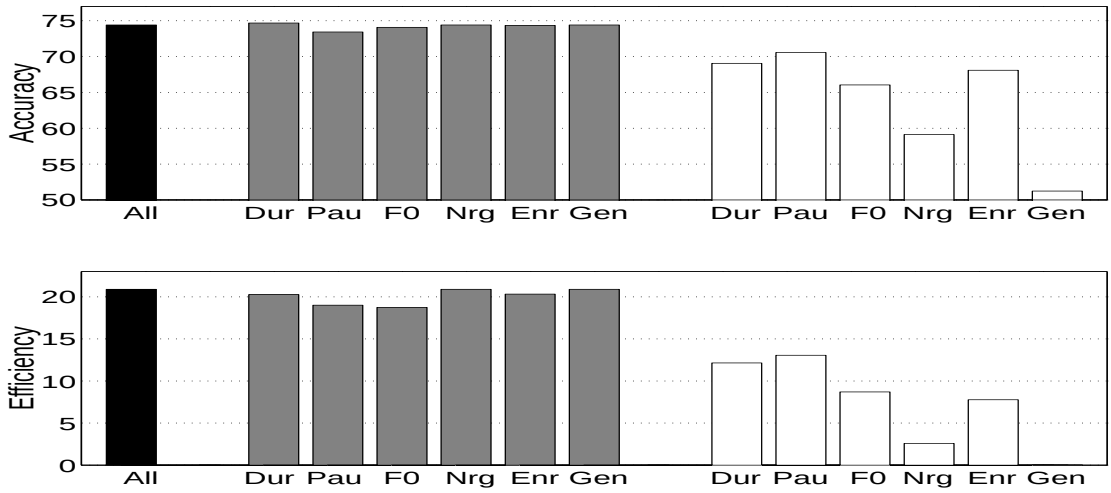

Figure 4: Performance of prosodic trees using different feature sets for the classification of statements and questions. N for each class=926. Chance accuracy = 50%. Gray bars=exclude feature type; white bars=include only feature type. Dur=Duration, Pau=Pause, F0=Fundamental frequency, Nrg=Energy, Enr=Enrate (speaking rate), Gen=Gender features.

Inspection of the leave-one-in tree results in Figure 4 indicates, not surprisingly, that the feature types most useful in the all-features analyses (duration and pause) yield the highest accuracies for the leave-one-in analyses (all of which are significantly above chance, $p < .0001$). What is interesting however is that enrate, which was used only minimally in the all-features tree, allows classification at 68.09%, which is

better than that of the F0-only tree. Furthermore, the enrate-only classifier is a mere shrub: as shown in Figure 5, it splits only once, on an *unnormalized* feature that expresses simply the variability in enrate over the utterance. As noted in the Method section, enrate is expected to correlate with speaking rate, although for this work we were not able to investigate the nature of this relationship. However the result has interesting potential implications. Theoretically, it suggests that absolute speaking rate may be less important for DA classification than *variation* in speaking rate over an utterance; a theory of conversation should be able to account for the lower variability in questions than statements. For applications, results suggest that enrate (which runs quickly) could be used alone to help distinguish these two types of DAs in a system in which other feature types are not available.

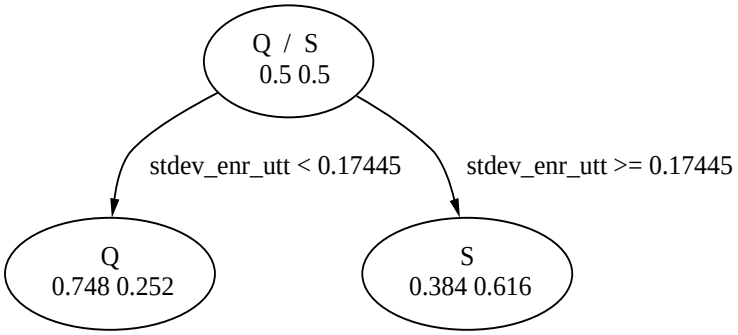

Figure 5: Decision tree for the classification of statements (S) and questions (Q), using only enrate features

We ran one further analysis on question classification. The aim was to determine the extent to which our grouping of different kinds of questions into one class affected the features used in question classification. As described in the Method section, our question class included yes-no questions, wh-questions, and declarative questions. Yet these different types of questions are expected to differ in their intonational characteristics (Haan et al., 1997a; van Heuven et al., 1997). Yes-no questions and declarative questions typically involve a final F0 rise; particularly declarative questions whose function is not conveyed syntactically. Wh-questions, on the other hand, often fall in F0, as do statements.

We broke our question class down into the originally-coded yes-no questions, wh-questions and declar-

ative questions, and ran a four-way classification along with statements. The resulting all-features tree is shown in Figure 6, and a summary of the feature usage is provided in Table 12.

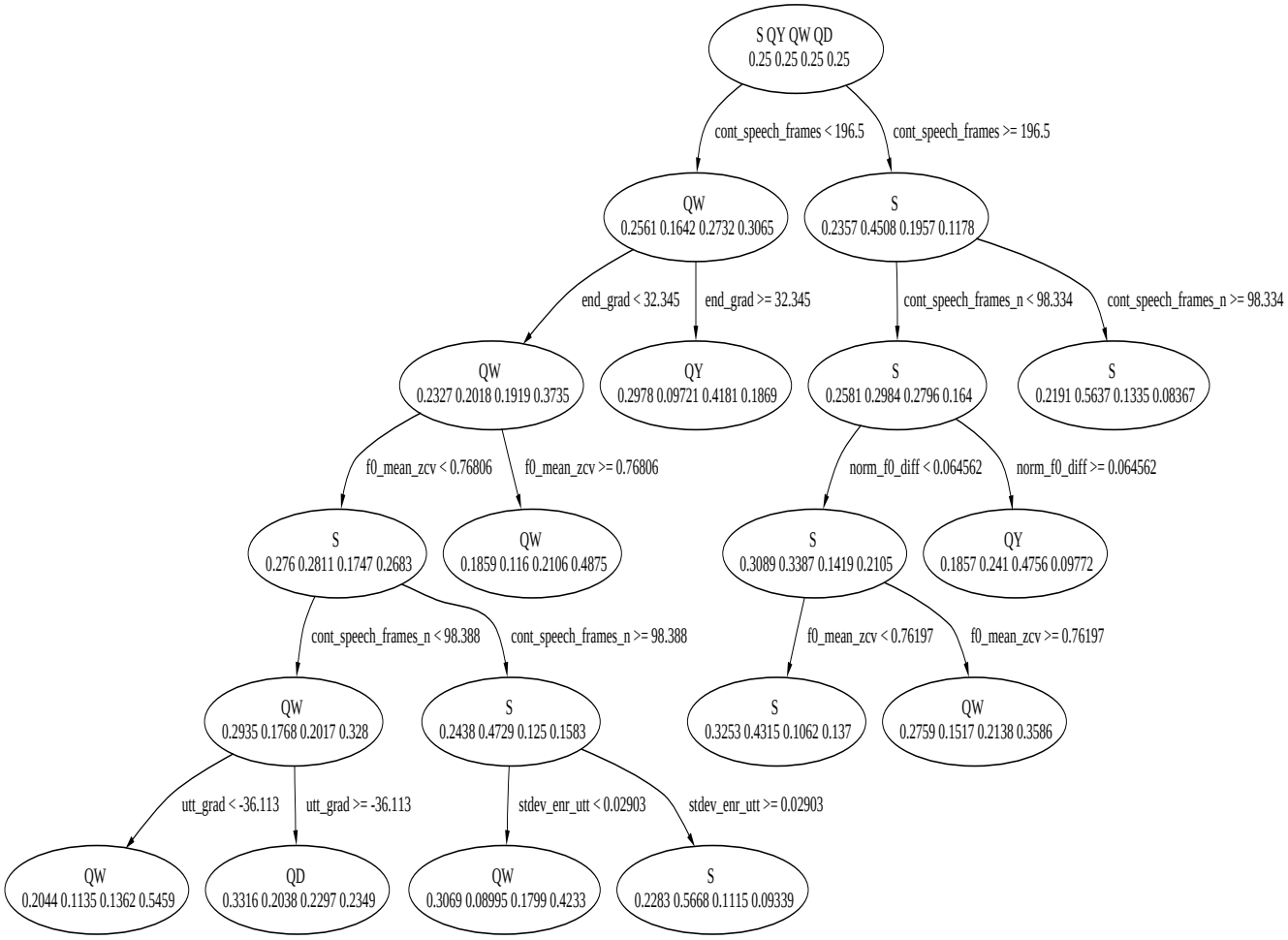

Figure 6: Decision tree for the classification of statements (S), yes-no questions (QY), wh-questions (QW), and declarative questions (QD)

Table 12: Feature Usage for Main Feature Types in Classification of Yes-No Questions, Wh-Questions, Declarative Questions, and Statements

Feature Type	Usage (%)
F0	0.432
Dur	0.318
Pau	0.213
Enr	0.037

The tree achieves an accuracy of 47.15%, a highly significant increase over chance accuracy (25%) by a binomial test, $p < .0001$. Unlike the case for the grouped question class, the most queried feature type is now F0. Inspection of the tree reveals that the pattern of results is consistent with the literature on question intonation. Final rises (end_grad, norm_f0_diff, and utt_grad) are associated with yes-no and declarative questions, but not with wh-questions. Wh-questions show a higher average F0 (f0_mean_zcv) than statements.

To further assess feature importance, we again built trees after selectively removing feature types. Results are shown in Figure 7.

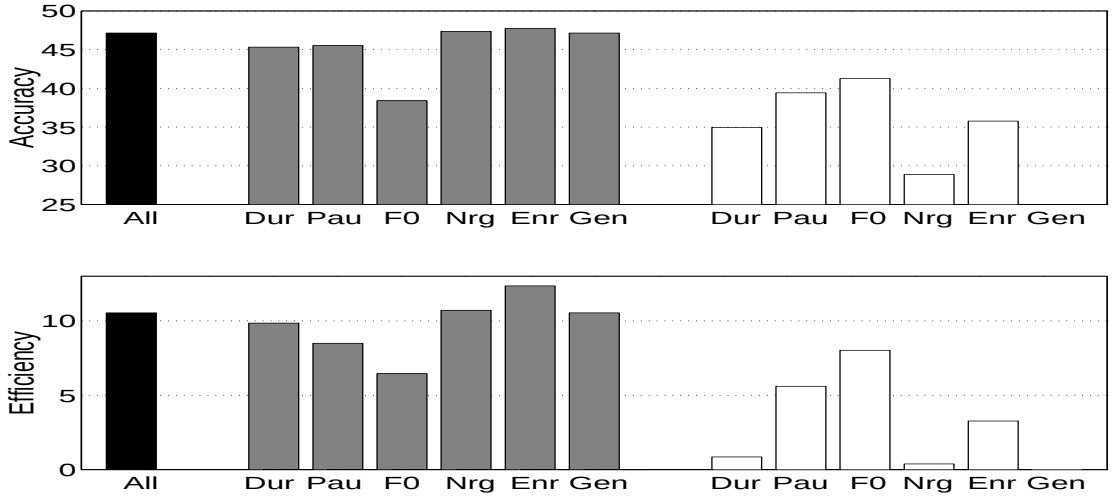

Figure 7: Performance of prosodic trees using different feature sets for the classification of Statements, Yes-No Questions, Wh-Questions, and Declarative Questions. N for each class=123. Chance=25%. Gray bars=exclude feature type; white bars=include only feature type. Dur=Duration, Pau=Pause, F0=Fundamental frequency, Nrg=Energy, Enr=Speaking rate, Gen=Gender features.

In contrast to Figure 4, in which accuracy was unchanged by removal of any single feature type, the present data show a sharp reduction in accuracy when F0 features are removed. This result is highly significant by a Sign test ($p < .001$, two-tailed) despite the small amount of data in the analyses due to downsampling to the size of the least frequent question subclass. For all other feature types, there was no significant reduction in accuracy when the feature type was removed. Thus, F0 plays an important role in question detection, but because different kinds of questions are signalled in different ways intonationally, combining questions into a single class as in the earlier analysis smooths over some of the distinctions. In particular, the grouping tends to conceal the features associated with the final F0 rise (probably because the rise is averaged in with final falls).

Integration with language model. To answer the question of whether prosody can aid question classification when word information is also available, tree probabilities were combined with likelihoods from our DA-specific trigram language models, using an optimal weighting factor. Results were computed for

the two test sets (HLD and DEV), and within the DEV set for both transcribed and recognized words. The outcome is shown in Table 13.

Table 13: Accuracy of Individual and Combined Models for the Classification of Questions

Knowledge Source	HLD Set true words	DEV Set true words	DEV Set N-best output
samples	1852	266	266
chance (%)	50.00	50.00	50.00
tree (%)	74.21	75.97	75.97
words (%)	83.65	85.85	75.43
words+tree (%)	85.64	87.58	79.76

The prosodic tree model yielded accuracies significantly better than chance for both test sets ($p < .0001$). The tree alone was also in fact more accurate than the recognized words alone for this task. Integration yielded consistent improvement over the words alone. The larger HLD set showed a highly significant gain in accuracy for the combined over the words-only model, $p < .001$ by a Sign test. Significance tests were not meaningful for the DEV set because of a lack of power given the small sample size; however the same pattern of results for the two sets is quite similar (in fact the spread is greatest for the recognized words), and therefore suggestive.

Subtask 2: Detection of incomplete utterances

A second problem area in the words-only analyses was the classification of incomplete utterances. Utterances labeled as incomplete in our work included three different main phenomena:[10]

Turn exits:		(A)	We have young children.
	→	(A)	So . . .
		(B)	Yeah, that's tough then.
Other-interruptions:	→	(A)	We eventually —
		(B)	Well you've got to start somewhere.
Self-interruptions:	→	(A)	And they were definitely —
(repairs)		(A)	At halftime they were up by four.

[10] In addition the class included a variety of utterance types deemed "uninterpretable" due to premature cut-off.

Although the three cases represent different phenomena, they are similar in that in each case the utterance could have been completed (and coded as the relevant type) but was not. An all-features tree built for the classification of incomplete utterances and all other classes combined ("non-incomplete") yielded an accuracy of 72.16% on the HLD test set, a highly significant improvement over chance, $p < .0001$.

Feature analyses. The all-features tree is complex and thus not shown, but feature usage by feature type is summarized in Table 14.

Table 14: Feature Usage for Main Feature Types in Classification of Incomplete Utterances and Non-Incomplete Utterances

Feature Type	Usage (%)
Dur	0.557
Nrg	0.182
Enr	0.130
F0	0.087
Pau	0.044

As indicated, the most-queried feature for this analysis is duration; not surprisingly, incomplete utterances are shorter overall than complete ones; certainly they are by definition shorter than their completed counterparts. Duration cannot however completely differentiate incomplete from non-incomplete utterances, because inherently short DAs (e.g. backchannels, agreements) are also present in the data. For these cases, other features such as energy and enrate play a role.

Results for trees run after selectively leaving out features are shown in Figure 8. Removal of duration features resulted in a significant loss in accuracy, 68.63%, $p < .0001$. Removal of any of the other feature types, however, did not significantly affect performance. Furthermore, a tree built using only duration features yielded an accuracy of 71.28%, which was not signficantly less accurate than the all-features tree. These results clearly indicate that duration features are primary for this task. Nevertheless, good accuracy could be achieved using other features types alone; for all other trees but the gender-only tree, accuracy was significantly above chance, $p < .0001$. Particularly noteworthy is the energy-only tree, which achieved an

accuracy of 68.97%. Typically, utterances fall to a low energy value when close to completion. However when speakers stop mid-stream, this fall has not yet occurred, and thus the energy stays unusually high. Inspection of the energy-only tree revealed that over 75% of the queries involved SNR rather than RMS features, suggesting that it is important to use features that can capture the energy from the speaker over the noise floor.

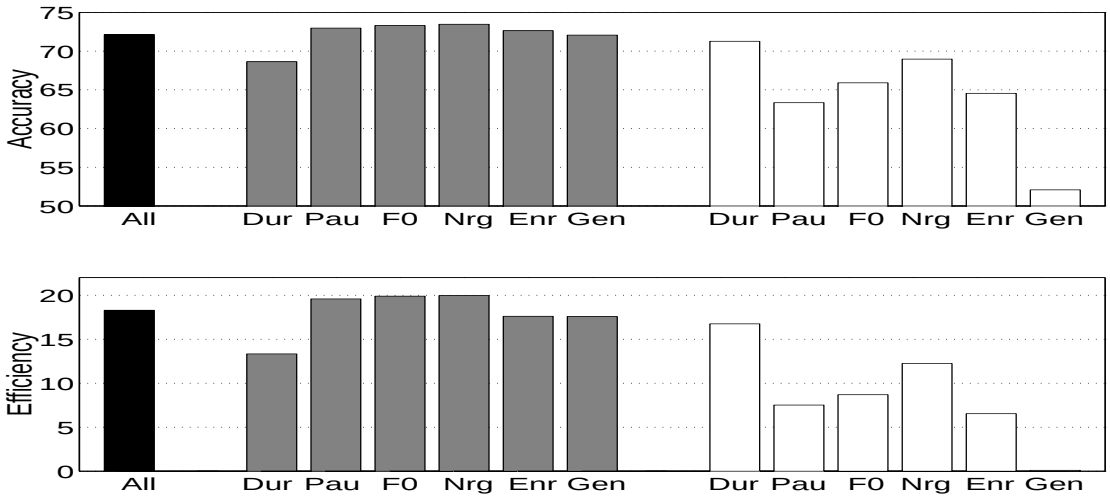

Figure 8: Performance of prosodic trees using different feature sets for the detection of Incomplete utterances from all other types. N for each class=1323. Chance=50%. Gray bars=exclude feature type; white bars=include only feature type. Dur=Duration, Pau=Pause, F0=Fundamental frequency, Nrg=Energy, Enr=Speaking rate, Gen=Gender features.

Integration with language model. We again integrated the all-features tree with a DA-specific language model to determine whether prosody could aid classification with word information present. Results are presented in Table 15. Like the earlier analyses, integration improves performance over the words-only model for all three test cases. Unlike earlier analyses, however, the relative improvement when true words are used is minimal, and the effect is not significant for either the HLD-true or the DEV-true data. However, the relative improvement for the DEV-N-best case is much larger. The effect is just below threshold for significance for this small dataset ($p = .067$), but would be expected based on the pattern results in the

previous analyses to easily reach significance for a set of data the size of the HLD set.

Table 15: Accuracy of Individual and Combined Models for the Classification of Incomplete Utterances

Knowledge Source	HLD Set true words	DEV Set true words	DEV Set N-best output
samples	2646	366	366
chance (%)	50.00	50.00	50.00
tree (%)	72.16	72.01	72.01
words (%)	88.44	89.91	82.38
words+tree (%)	88.74	90.49	84.56

Results suggest that for this task, prosody is an important knowledge source when word recognition is not perfect. When true words are available however, it is not clear whether adding prosody aids performance.

One factor underlying this pattern of results may be that the tree information is already accounted for in the language model. Consistent with this possibility is the fact that the tree uses mainly duration features, which are indirectly represented in the language model by the end-of-sentence marker. On the other hand, typically the word length of true and N-best lists are similar, and our results differ for the two cases, so it is unlikely that this could be the other factor.

Another possibilitiy is that when true words are available, certain canonical incomplete utterances can be detected with excellent accuracy. A likely candidate here is the turn exit. Turn exits typically contain one or two words from a small inventory of possibilities—mainly coordinating conjunctions (“and”, “but”) and fillers (“uh”, “um”). Similarly, because Switchboard consists mainly of first-person narratives, a typical self-interrupted utterance in this corpus is a noncommittal false start such as “I—” or “I think—”. Both the turn exits and the noncommittal false starts are lexically cued and are thus likely to be well captured by a language model that has true words available.

A third possible reason for the lack of improvement over true words is that the prosodic model loses sensitivity because it averages over phenomena with different characteristics. False starts in our data typically involved a sudden cut-off, whereas for turn exits the preceding speech was often drawn out as in a hesitation. As a preliminary means of investigating this possibility, we built a tree for incomplete utterances only, but

breaking down the class into those ending at turn boundaries (mainly turn exits and interrupted utterances) versus those ending within a speaker's turn (mainly false starts). The resulting tree achieved high accuracy (81.17%), and revealed that the two subclasses differed on a number of features. For example, false starts were longer in duration, higher in energy, and had faster speaking rates than the turn exit/other-interrupted class. These features were less useful in the original task. Thus, as we also saw for the case of question detection, a prosodic model for incomplete utterances is probably best built on data that has been broken down to isolate subsets of phenomena whose prosodic features pattern differently.

Subtask 3: Detection of agreements

Our final subtask examined whether prosody could aid in the classification of classifying explicit agreements (e.g., "that's exactly right"). As shown earlier, agreements were most often misclassified as backchannels (e.g. "uh-huh", "yeah"). Thus our experiments focussed on the distinction by including only these two DAs in the trees. An all-features tree for this task classified the data with an accuracy of 68.77% (significantly above chance by a binomial test, $p < .0001$), and with an efficiency of 12.21%

Feature analyses. The all-features tree is shown in Figure 9. It uses duration, pause, and energy features. From inspection we see that agreements are consistently longer in duration and have higher energy (as measure by mean SNR) than backchannels. The pause feature in this case may play a role similar to that discussed for the question classification task. Although agreements and backchannels were about equally likely to occur turn-finally, backchannels were more than three times as likely as agreements to be the only DA in a turn. Thus backchannels were more often surrounded by nonspeech regions (pauses during which the other speaker was typically talking), causing the cont_speech_frames window to not be filled at the edges of the DA and thereby lowering the value of the feature.

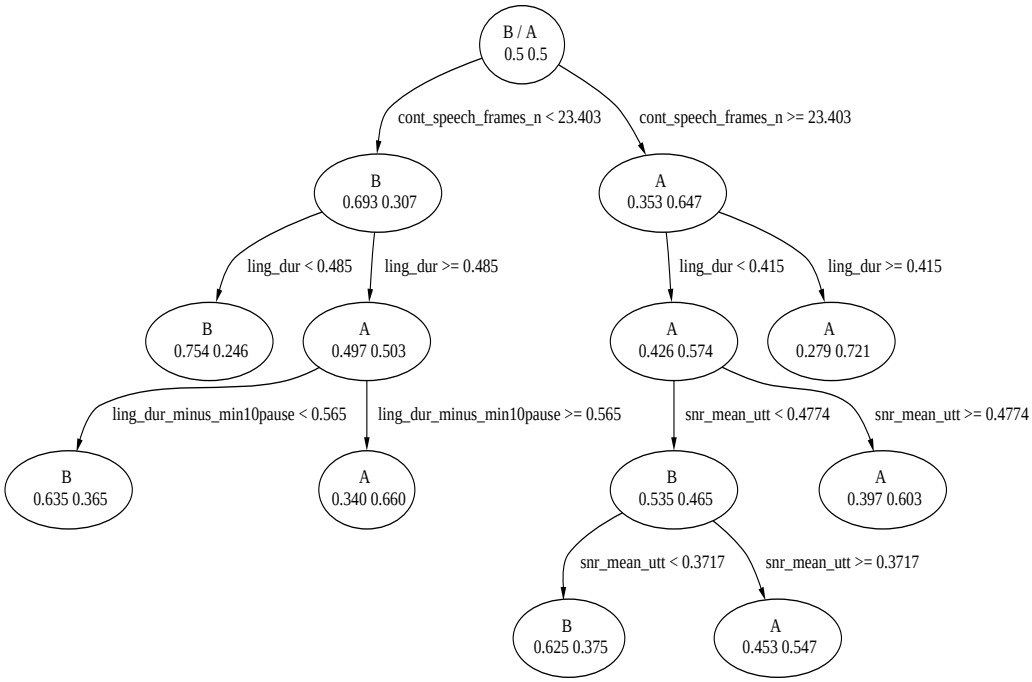

Figure 9: Decision tree for the classification of Backchannels (B) and Agreements (A).

Significance tests for the leave-one-out trees showed that removal of the main feature types used in the all-features tree, i.e. duration, pause, and energy features resulted in a significant reduction in classification accuracy, $p < .001$, $p < .05$, and $p < .05$, respectively. Although significant, the reduction was not large in absolute terms, as seen from the figure and the α levels for significance. For the leave-one-in trees, results were in all cases significantly lower than that of the all-features trees; however duration and pause features alone each yielded accuracy rates near that of the all-features tree. Although neither F0 nor enrate were used in the all-features tree, each alone was able to distinguish the DAs at rates significantly better than chance ($p < .0001$).

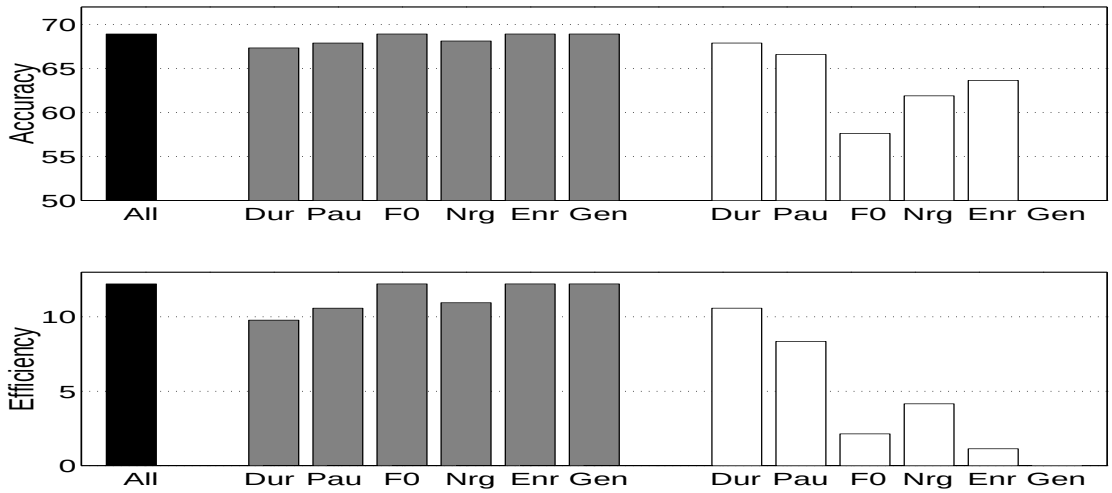

Figure 10: Performance of prosodic trees using different feature sets for the classification of Backchannels and Agreements. N for each class=1260. Chance=50%. Gray bars=exclude feature type; white bars=include only feature type. Dur=Duration, Pau=Pause, F0=Fundamental frequency, Nrg=Energy, Enr=Speaking rate, Gen=Gender features.

Integration with language model. Integration results are reported in Table 16. A number of observations are noteworthy. First, integrating the tree with word models improves performance considerably for all three test sets. Sign tests run for the larger HLD set showed a highly significant gain in accuracy by adding prosody, $p < .00001$; the DEV set did not contain enough samples for sufficient power to reject the null hypothesis, but show the same pattern of results and thus would be expected to reach significance for a larger data set. Second, for this analysis, the prosodic tree has better accuracy than the true words for the HLD set. Third, comparison of the data for the different test sets reveals an unusual pattern of results. Typically (and in the previous analyses), accuracy results for tree and word models were better for the HLD than for the DEV set. As noted in the Method section, HLD waveforms were segmented into DAs in the same manner (automatically) as the training data, while DEV data was carefully segmented by hand. For this task however, results for both tree and word models are considerably better for the DEV data, i.e., the mismatched case. This pattern can be understood as follows. In the automatically segmented training and

HLD data, utterances with "bad" estimated start or end times were thrown out of the analysis as described in the Method section. The DAs most affected by the bad time marks were very short DAs, many of which were brief, single-word backchannels such as "yeah". Thus the data remaining in the training and HLD sets are biased toward longer DAs, while the data in the DEV set retain the very brief DAs. Since the present task pits backchannels against the longer agreements, an increase in the percentage of shorter backchannels (from training to test, as occurs when testing on the DEV data) should only enhance discriminability for the prosodic trees as well as for the language model.

Table 16: Accuracy of Individual and Combined Models for the Classification of Agreements

Knowledge Source	HLD Set true words	DEV Set true words	DEV Set N-best output
samples	2520	214	214
chance (%)	50.00	50.00	50.00
tree (%)	68.77	72.88	72.88
words (%)	68.63	80.99	78.22
words+tree (%)	76.90	84.74	81.70

SUMMARY AND GENERAL DISCUSSION

Feature importance

Across analyses we found that a variety of features were useful for DA classification. Results from the leave-one-out and leave-one-in trees showed that there is considerable redundancy in the features; typically there is little loss when one feature type is removed. Interestingly, although canonical or predicted features such as F0 for questions are important, less predictable features (such as pause features for questions) actually show similar or even greater influence on results.

Duration was found to be important not only in the seven-way classification, which included both long and short utterance types, but also for subtasks within general length categories (e.g., statements versus questions, backchannels versus agreements). Duration was also found to be useful as an added knowledge

source to language model information, even though the length in words of an utterance is indirectly captured by the language model. Across tasks, the most-queried duration features were not raw duration, but rather duration-related measures that relied on the computation of other feature types.

F0 information was found to be important, as expected, for the classification of questions, particularly when questions were broken down by type. However it was also of use in many other classification tasks. In general, the main contribution from F0 features for all but the question task came from global features (such as overall mean or gradient) rather than local features (such as the penultimate and end features, or the intonational event features). An interesting issue to explore in future work is whether this is a robustness effect, or whether global features are inherently better predictors of DAs than local features such as accents and boundaries.

Energy features were particularly helpful for classifying incomplete utterances, but also for the classification of agreements and backchannels. Analysis of the usage of energy features over all tasks revealed that features based on the signal-to-noise ratio were queried more than 4.8 times as often as features based on the RMS energy. Similarly, when the individual leave-one-in analyses for energy features were computed using only RMS versus only SNR features, results were consistently better for the SNR experiments. This suggests that for speech data collected under noisy conditions, it is worthwhile to estimate the energy of the speaker above the noise floor.

Enrate, the experimental speaking-rate feature from ICSI, proved to be useful across analyses in the following way. Although no task was significantly effected when enrate features were removed, enrate systematically achieved good performance when used alone. It was always better alone than at least one of the other main feature types alone (excluding gender). Furthermore it provided remarkable accuracy for the classification of questions and statements, without any conversation-level normalization. Thus the measure could be a valuable feature to include in a system, particularly if other more costly features cannot be computed.

Finally, across analyses, gender was not used in the trees. This suggests that gender-dependent features such as F0 were sufficiently normalized to allow gender-independent modeling. Since many of the features

were normalized with respect to all values from a conversation side, it is possible that men and women do differ in the degree to which they use different prosodic features (even after normalization for pitch range), but that we cannot discern these differences here because speakers have been normalized individually.

Overall, the high degree of feature compensation found across tasks suggests that automatic systems could be successful using only a subset of the feature types. However we also found that different feature types are used to varying degrees in the different tasks, and it is not straightforward at this point to predict which features will be most important for a task. Therefore, for best coverage on a variety of classification tasks, it is desirable to have as many different feature types available as possible.

Integration of trees with language models

Not only were the prosodic trees able to classify the data at rates significantly above chance, but they also consistently provided a advantage over word information alone. To summarize the integration experiments: all tasks with the exception of the incomplete-utterance task showed a significant improvement over words alone for the HLD set. For the incomplete utterance task, results for the DEV set were marginally significant. In all cases, the DEV set lacked power due to small sample size, making it difficult to reach significance in the comparisons. However the relative win on the DEV set was consistently larger for the experiments using recognized rather than true words. This pattern of results suggests that prosody can provide significant benefit over word information alone, particularly when word recognition is imperfect.

FUTURE WORK

One aim of future work in this area is to optimize the prosodic features, and better understand the correlations among them. In evaluating the contribution of features, it is important to take into account such factors as measurement robustness and inherent constraints leading to missing data in our trees. For example, duration is used frequently, but it is also (unlike, for example, F0 information) available and fairly accurately extracted for all utterances. We would also like to better understand which of our features capture functional versus semantic or paralinguistic information, as well as the extent to which features are

speaker-dependent.

A second goal is to explore additional features that do not depend on the words. For example, we found that whether or not an utterance is turn initial and/or turn final, and the rate of interruption (including overlaps) by the other speaker can significantly improve tree performance for certain tasks. In our overall model, we consider turn-related features to be part of the dialog grammar. Nevertheless, if one wanted to design a system that did not use word information, turn features could be used along with the prosodic features to improve performance overall.

Third, although we chose to use decision trees for the reasons given earlier, we might have used any suitable probabilistic classifier, i.e., any model that estimates the posterior probabilities of DAs given the prosodic features. We have conducted preliminary experiments to assess how neural networks compare to decision trees for the type of data studied here. Neural networks are worth investigating since they offer a number of potential advantages over decision trees. They can learn decision surfaces that lie at an angle to the axes of the input feature space, unlike standard CART trees which always split continuous features on one dimension at a time. The response function of neural networks is continuous (smooth) at the decision boundaries, allowing them to avoid hard decisions and the complete fragmentation of data associated with decision tree questions. And most importantly, neural networks with hidden units can learn new features that combine multiple input features. Results from preliminary experiments on a single task showed that a softmax network (Bridle, 1990) without hidden units resulted in a slight improvement over a decision tree on the same task. The fact that hidden units did not afford an advantage indicates that complex combinations of features (as far as the network could learn them) do not better predict DAs for the task than linear combinations of our input features.

Thus, whether or not substantial gains can be obtained using alternative classifier architectures remains an open question for future work. One approach that looks promising given the redundancy among different feature types is a combination of parallel classifiers, each based on subcategory of features, e.g., using the mixture-of-experts framework (Jordan and Jacobs, 1993). We will also need to develop an effective way to combine specialized classifiers (such as those investigated for the subtasks in this study) into an overall

classifier for the entire DA set.

Finally, many questions remain concerning the best way to integrate the various knowledge sources. Instead of treating words and prosody as independent knowledge sources, as done here for simplicity, we could provide both types of cues to a single classifier. This would allow the model to account for interactions between prosodic cues and words, such as word-specific prosodic patterns. The main problem is the large number of potential input values that "word features" can take on. A related question is how to combine prosodic classifiers most effectively with dialog grammars and the contextual knowledge sources.

CONCLUSION

We have shown that in a large database of truly natural human-human conversations, assuming equal class prior probabilities, prosody is a useful knowledge source for a variety of DA classification tasks. The features that allow this classification are task-dependent. Although canonical features are used in predicted ways, other less obvious features also play important roles. Overall there is a high degree of correlation among features, such that if one feature type is not available, others can compensate. Finally, integrating prosodic decision trees with DA-specific statistical language models improves performance over that of the language models alone. We conclude that DAs are redundantly marked in free conversation, and that a variety of automatically extractable prosodic features could aid the processing of natural dialog in speech applications.

References

Allen, J. and M. Core (1997). Draft of DAMSL: Dialog act markup in several layers.

Batliner, A., C. Weiand, A. Kießling, and E. Nöth (1993). Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody. In *Working Papers 41, Proc. ESCA Workshop on Prosody*, Lund, Sweden, pp. 112–115.

Breiman, L., J. H. Friedman, R. A. Olshenn, and C. J. Stone (1983). *Classification and Regression Trees*. Pacific Grove, California: Wadsworth & Brooks.

Bridle, J. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Soulie and J.Herault (Eds.), *Neurocomputing: Algorithms, Architectures and Applications*, pp. 227–236. Berlin: Springer.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics 22*(2), 249–254.

Clark, H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Geluykens, R. and M. Swerts (1993). Local and global prosodic cues to discourse organization in dialogues. In *Working Papers 41, Proc. ESCA Workshop on Prosody*, Lund, Sweden, pp. 108–111.

Godfrey, J., E. Holliman, and J. McDaniel (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92*, San Francisco, pp. 517–520.

Grosz, B. and J. Hirschberg (1992). Some intonational characteristics of discourse structure. In *Proc. ICSLP*, Volume 1, Banff, Canada, pp. 429–432.

Haan, J., V. J. van Heuven, J. J. A. Pacilly, and R. van Bezooijen (1997a). An anatomy of dutch question intonation. In H. de Hoop and M. den Dikken (Eds.), *Linguistics in the Netherlands*. Amsterdam: John Benjamins.

Haan, J., V. J. van Heuven, J. J. A. Pacilly, and R. van Bezooijen (1997b). Intonational characteristics of declarativity and interrogativity in Dutch: A comparison. In A. Botonis, G. Kouroupetroglou, and

G. Carayiannis (Eds.), *ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, Greece, pp. 173–176.

Hirschberg, J. and C. Nakatani (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of ACL-96*, pp. 286–293.

Jordan, M. I. and R. A. Jacobs (1993). Hierarchical mixtures of expert and the EM algorithm. Memo 1440, Artifcial Intelligence Laboratory, MIT, Cambridge, Mass.

Jurafsky, D., R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema (1997a). Switchboard discourse language modeling project report. Technical report, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD.

Jurafsky, D., R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema (1997b). Automatic detection of discourse structure for speech recognition and understanding. In *IEEE Workshop on Speech Recognition and Understanding*, Santa Barbara, CA.

Jurafsky, D., E. Shriberg, and D. Biasca (1997). Switchboard-DAMSL Labeling Project Coder's Manual. http://stripe.colorado.edu/~jurafsky/manual.august1.html.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Trans. Acoustics, Speech and Signal Processing 35*(3), 400–401.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Mast, M., R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, and V. Warnke (1996). Dialog act classification with the help of prosody. In *ICSLP-96*, Volume 3, Philadelphia, pp. 1732–1735.

Meteer, M. et al. (1995). *Dysfluency Annotation Stylebook for the Switchboard Corpus*. Linguistic Data Consortium. Revised June 1995 by Ann Taylor. ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps.gz.

Morgan, N., E. Fosler, and N. Mirghafori (1997). Speech recognition using on-line estimation of speaking rate. In *EUROSPEECH-97*, Volume 4, Rhodes, Greece, pp. 2079–2082.

Neumeyer, L. and M. Weintraub (1994). Microphone-independent robust signal processing using probabilistic optimum filtering. In *ARPA HLT Workshop*, Plainsboro, NJ, pp. 336–341.

Neumeyer, L. and M. Weintraub (1995). Robust speech recognition in noise using adaptation and mapping techniques. In *ICASSP-95*, Volume 1, Detroit, pp. 141–144.

Pierrehumbert, J. and J. Hirschberg (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack (Eds.), *Intentions in Communication*, pp. 271–311. Cambridge, Mass.: MIT Press.

Shriberg, E., R. Bates, and A. Stolcke (1997). A prosody-only decision-tree model for disfluency detection. In *EUROSPEECH-97*, Volume 5, Rhodes, Greece, pp. 2383–2386.

Siegel, S. and N. J. Castellan, Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences* (Second ed.). New York: McGraw-Hill.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America 101*, 514–521.

Swerts, M. and M. Ostendorf (1997). Prosodic indications of discourse structure in human-machine interactions. *Speech Communication 22*(1).

Taylor, P., S. King, S. Isard, H. Wright, and J. Kowtko (1997a). Using intonation to constrain language models in speech recognition. In *EUROSPEECH-97*, Rhodes, Greece. to appear.

Taylor, P. A., S. King, S. D. Isard, H. Wright, and J. Kowtko (1997b). Using intonation to constrain language models in speech recognition. In *EUROSPEECH-97*, Volume 5, pp. 2763–2766.

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler and D. R. Ladd (Eds.), *Prosody: Models and Measurements*, Chapter 5, pp. 53–66. Berlin: Springer Verlag.

van Heuven, V. J., J. Haan, E. Janse, and E. J. van der Torre (1997). Perceptual identification of sentence type and the time-distribution of prosodic interrogativity markers in Dutch. In A. Botonis, G. Kouroupetroglou, and G. Carayiannis (Eds.), *ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, Greece, pp. 317–320.

Warnke, V., R. Kompe, H. Niemann, and E. Nöth (1997). Integrated dialog act segmentation and classification using prosodic features and language models. In *Proc. EUROSPEECH-97*, Volume 1, Rhodes, Greece, pp. 207–210.

APPENDIX: TABLE OF ORIGINAL DIALOG ACTS

The following table lists the 42 original (before grouping into classes) dialog acts. Counts and relative frequencies were obtained from the corpus of 197,000 utterances used in model training.

Tag	Abbrev	Example	Cnt	%
Statement-non-opinion	sd	*Me, I'm in the legal department.*	72,824	36%
Acknowledge (Backchannel)	b	*Uh-huh.*	37,096	19%
Statement-opinion	sv	*I think it's great*	25,197	13%
Agree/Accept	aa	*That's exactly it.*	10,820	5%
Abandoned or Turn-Exit	% . . . -/	*So, -/*	10,569	5%
Appreciation	ba	*I can imagine.*	4,633	2%
Yes-No-Question	qy	*Do you have to have any special training?*	4,624	2%
Non-verbal	x	*<Laughter>,<Throat_clearing>*	3,548	2%
Yes answers	ny	*Yes.*	2,934	1%
Conventional-closing	fc	*Well, it's been nice talking to you.*	2,486	1%
Uninterpretable	%	*But, uh, yeah*	2,158	1%
Wh-Question	qw	*Well, how old are you?*	1,911	1%
No answers	nn	*No.*	1,340	1%
Response Acknowledgement	bk	*Oh, okay.*	1,277	1%
Hedge	h	*I don't know if I'm making any sense or not.*	1,182	1%
Declarative Yes-No-Question	qy^d	*So you can afford to get a house?*	1,174	1%
Other	o,fo	*Well give me a break, you know.*	1,074	1%
Backchannel in question form	bh	*Is that right?*	1,019	1%
Quotation	^q	*You can't be pregnant and have cats*	934	.5%
Summarize/reformulate	bf	*Oh, you mean you switched schools for the kids.*	919	.5%
Affirmative non-yes answers	na	*It is.*	836	.4%
Action-directive	ad	*Why don't you go first*	719	.4%
Collaborative Completion	^2	*Who aren't contributing.*	699	.4%
Repeat-phrase	b^m	*Oh, fajitas*	660	.3%
Open-Question	qo	*How about you?*	632	.3%
Rhetorical-Questions	qh	*Who would steal a newspaper?*	557	.2%
Hold before answer/agreement	^h	*I'm drawing a blank.*	540	.3%
Reject	ar	*Well, no*	338	.2%
Negative non-no answers	ng	*Uh, not a whole lot.*	292	.1%
Signal-non-understanding	br	*Excuse me?*	288	.1%
Other answers	no	*I don't know*	279	.1%
Conventional-opening	fp	*How are you?*	220	.1%
Or-Clause	qrr	*or is it more of a company?*	207	.1%
Dispreferred answers	arp,nd	*Well, not so much that.*	205	.1%
3rd-party-talk	t3	*My goodness, Diane, get down from there.*	115	.1%
Offers, Options & Commits	oo,cc,co	*I'll have to check that out*	109	.1%
Self-talk	t1	*What's the word I'm looking for*	102	.1%
Downplayer	bd	*That's all right.*	100	.1%
Maybe/Accept-part	aap/am	*Something like that*	98	<.1%
Tag-Question	^ g	*Right?*	93	<.1%
Declarative Wh-Question	qw^d	*You are what kind of buff?*	80	<.1%
Apology	fa	*I'm sorry.*	76	<.1%
Thanking	ft	*Hey thanks a lot*	67	<.1%

List of Figures

List of Tables