PROSODY MODELING FOR AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING

ELIZABETH SHRIBERG* AND ANDREAS STOLCKE*

Abstract. This paper summarizes statistical modeling approaches for the use of prosody (the rhythm and melody of speech) in automatic recognition and understanding of speech. We outline effective prosodic feature extraction, model architectures, and techniques to combine prosodic with lexical (word-based) information. We then survey a number of applications of the framework, and give results for automatic sentence segmentation and disfluency detection, topic segmentation, dialog act labeling, and word recognition.

Key words. Prosody, speech recognition and understanding, hidden Markov models.

1. Introduction. Prosody has long been studied as an important knowledge source for speech understanding. In recent years there has been a large amount of computational work aimed at prosodic modeling for automatic speech recognition and understanding.[1] Whereas most current approaches to speech processing model only the words, prosody provides an additional knowledge source that is inherent in, and exclusive to, spoken language. It can therefore provide additional information that is not directly available from text alone, and also serves as a partially redundant knowledge source that may help overcome the errors resulting from faulty word recognition.

In this paper, we summarize recent work at SRI International in the area of computational prosody modeling, and results from several recognition tasks where prosodic knowledge proved to be of help. We present only a high-level perspective and summary of our research; for details the reader is referred to publications cited.

2. Modeling philosophy. Most problems for which prosody is a plausible knowledge source can be cast as statistical classification problems. By that we mean that some linguistic unit U (e.g., words or utterances) is to be classified as one of several target classes S. The role of prosody is to provide us with a set of *features* F that can help predict S. In a probabilistic framework, we wish to estimate $P(S|F)$. In most such tasks it is also a good idea to use the information contained in the *word sequence* W associated with U, and we therefore generalize the modeling task to estimate $P(S|W, F)$. In fact, W and F are not restricted to

*SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, Email {ees,stolcke}@speech.sri.com. We thank our many colleagues at SRI, ICSI, University of Washington (formerly at Boston University), and the 1997 Johns Hopkins CLSP Summer Workshop, who were instrumental in much of the work reported here. The research was supported by NSF Grants IRI-9314967, IRI-9618926, and IRI-9619921, by DARPA contract no. N66001-97-C-8544, and by NASA contract no. NCC 2-1256. Additional support came from the sponsors of the 1997 CLSP Workshop [7, 11] and from the DARPA Communicator project at UW and ICSI [8]. The views herein are those of the authors and should not be interpreted as representing the policies of the funding agencies.

[1]Too much work in fact, to cite here without unfair omissions. We cite some specifically relevant work below; a more comprehensive list can be found in the papers cited.

pertain only to the unit in question; they may refer to the context of U as well. For example, when classifying an utterance into dialog acts, it is important to take the surrounding utterances into account.

Starting from this general framework, and given a certain labeling task, many decisions must be made to use prosodic information effectively. What is the nature of the features F to be used? How can we model the relationship between F and the target classes S? How should we model the effect of lexical information W and its interaction with prosodic properties F? In the remainder of this paper we give a general overview of approaches that have proven successful for a variety of tasks.

2.1. Direct modeling of target classes. A crucial aspect of our work, as well as that of some other researchers [6, 5] is that the dependence between prosodic features and target classes (e.g., dialog acts, phrase boundaries) is modeled directly in a statistical classifier—without the use of intermediate abstract phonological categories, such as pitch accent or boundary tone labels. This bypasses the need to hand-annotate such labels for training purposes, avoids problems of annotation reliability, and allows the model to choose the level of granularity of the representation that is best suited for the task [2].

2.2. Prosodic features. As predictors of the target classes, we extract features from a forced alignment of the transcripts (usually with phone-level alignment information), which can be based on either true words, or on (errorful) speech recognition output. Similar approaches are used by others [2]. This yields a rich inventory of "raw" features reflecting F0, pause and segment durations, and energy. From the raw features we compute a wide range of "derived" features—devised (we hope) to capture characteristics of the classes—which are normalized in various ways, conditioned on certain extraction regions, or conditioned on values of other features.

Phone-level alignments from a speech recognizer provide durations of pauses and various measures of lengthening (we have used syllable, rhyme, and vowel durations for various tasks) and speaking rate. Pitch-based features benefit greatly from a postprocessing stage that regularizes the raw F0 estimates and models octave errors [10]. As a byproduct of the postprocessing, we also obtain estimates of the speaker's F0 baseline, which we have found useful for pitch range normalizations.

Combined with F0 estimates, the recognizer output also allows computation of pitch movements and contours over the length of utterances or individual words, or over the length of windows positioned relative to a location of interest (e.g., around a word boundary). The same applies to energy-based features.

2.3. Prosodic models. Any number of statistical classifiers that can deal with a mix of categorical and real-valued features may be used to model $P(S|F, W)$. These requirements, as well as our desire to be able to inspect our models (both to understand patterns and for sanity checking), have led us to use mainly decision trees as classifiers. Decision trees have two main problems, how-

ever, which we have tried to address. First, to help overcome the problem of greediness, we wrap a feature subset selection algorithm around the standard tree growing algorithm, thereby often finding better classifiers by eliminating detrimental features up front from consideration by the tree [9]. Second, to make the trees sensitive to prosodic features in the case of highly skewed class sizes, we train on a resampled version of the target distribution in which all classes have equal prior probabilities. This approach has additional benefits. It allows prosodic classifiers to be compared (both qualitatively and quantitatively) across different corpora and tasks. In addition, classifiers based on uniform prior distributions are well suited for integration with language models, as described below.

2.4. Lexical models. Our target classes are typically cued by both lexical and prosodic information; we are therefore interested in optimal modeling and combination of both feature types. Although in principle one could add words directly as input features to a prosodic classifier, in practice this is often not feasible since it results in too large a feature space for most classifiers. Approaches for cardinality reduction (such as inferring word classes via unsupervised clustering [4]) offer promise and are an area we are interested in investigating. To date, however, we have used statistical language models (LMs) familiar from speech recognition. One or more LMs are used to effectively model the joint distribution of target classes S and words W, $P(W, S)$. With labeled training data, such models can usually be estimated in a straightforward manner. During testing on unlabeled data, we compute $P(S|W)$ to predict the possible classes and their posterior probabilities, or simply to recover the most likely target class given the words.

2.5. Model combination. The prosodic model may be combined with a language model in different ways, including

- *Posterior interpolation:* Compute $P(S|F, W)$ via the prosodic model and $P(S|W)$ via the language model and form a linear combination of the two. The weighting is optimized on held-out data. This is a weak combination approach that does not attempt to model a more fine-grained structural relationship between the knowledge sources, but it also does not make any strong assumptions about their independence.
- *Posteriors as features:* Compute $P(S|W)$ and use the LM posterior estimate as an additional feature in the prosodic classifier. This approach can capture some of the dependence between the knowledge sources. However, in practice it suffers from the fact that the LM posteriors on the training data are often strongly biased, and therefore lead the tree to over-rely on them unless extra held-out data is used for training.
- *HMM-based integration:* Compute likelihoods $P(F|S, W)$ from the prosody model and use them as observation likelihoods in a hidden Markov model (HMM) derived from the LM.[2] The HMM is constructed

[2] By equating the class distributions for classifier training, as advocated above, we obtain posterior estimates that are proportional to likelihoods, and can therefore be used directly in the HMM.

TABLE 1
Sentence boundary and disfluency event tagging error rates for the Switchboard corpus. The higher chance error rate for recognized words is due to incorrect word boundary hypotheses.

Model	True words	Recognized words
LM only	7.3	26.2
Prosody only	11.1	27.1
Combined	6.9	25.1
Chance	18.2	30.8

to encode the unobserved classes S in its state space. By associating these states with prosodic likelihoods we obtain a joint model of F, S, and W, and HMM algorithms can be used to compute the posteriors $P(S|F,W)$ that incorporate all available knowledge.

This approach models the relationship between words and prosody at a detailed level, but it does require the assumption that prosody and words are conditionally independent given the labels S. In practice, however, this model often works very well even if the independence assumption is clearly violated.

For a detailed discussion of these approaches, and results showing their relative success under various conditions, see [12, 9, 15].

3. Applications. Having given a brief overview of the key ideas in our approach to computational prosody, we now summarize some applications of the framework.

3.1. Sentence segmentation and disfluency detection. The framework outlined was applied to the detection of sentence boundaries and disfluency interruption points in both conversational speech (Switchboard) and Broadcast News [12, 9]. The target classes S in this case were labels at each word boundary identifying the type of event: sentence boundary, various types of disfluencies (e.g., hesitations, repetitions, deletions) and fluent sentence-internal boundaries. The prosodic model was based on features extracted around each word boundary, capturing pause and phone durations, F0 properties, and ancillary features such as whether a speaker change occurred at that location.

The LM for this task was a hidden event N-gram, i.e., an N-gram LM in which the boundary events were represented by tags occurring between the word tokens. The LM was trained like a standard N-gram model from tagged training text; it thus modeled the joint probability of tags and words. In testing, we ran the LM as an HMM in which the states correspond to the unobserved (hidden) boundary events. Prosodic likelihood scores $P(F|S,W)$ for the boundary events were attached to these states as described above, to condition the HMM tagging output on the prosodic features F.

We tested such a model for combined sentence segmentation and disfluency detection on conversational speech, where it gave about 7% boundary classification error using correct word transcripts. The results for various knowledge

TABLE 2
Sentence boundary tagging error rates for two different speech corpora: Switchboard (SWB) and Broadcast News (BN).

	SWB		BN	
Model	True words	Rec. words	True words	Rec. words
LM only	4.3	22.8	4.1	11.8
Prosody only	6.7	22.9	3.6	10.9
Combined	4.0	22.2	3.3	10.8
Chance	11.0	25.8	6.2	13.3

sources based on true and recognized words are summarized in Table 1 (adapted from [12]). For both test conditions, the prosodic model improves the accuracy of an LM-only classifier by about 4% relative.

We also carried out a comparative study of sentence segmentation alone, comparing Switchboard (SWB) telephone conversations to Broadcast News (BN) speech. Results are given in Table 2 (adapted from [9]). Again the combination of word and prosodic knowledge yielded the best results, with significant improvements over either knowledge source alone.

A striking result in BN segmentation was that the prosodic model alone performed better than the LM alone. This was true even when the LM was using the correct words, and even though it was trained on two orders of magnitude more data than the prosody model. Pause duration was universally the most useful feature for these tasks; in addition, SWB classifiers relied primarily on phone duration features, whereas BN classifiers made considerable use of pitch range features (mainly distance from the speaker's estimated baseline). We attribute the increased importance of pitch features in BN to the higher acoustic quality of the audio source, and the preponderance of professional speakers with a consistent speaking style.

3.2. Topic segmentation in Broadcast News. A second task we looked at was locating topic changes in a broadcast news stream, following the DARPA TDT [3] framework. For this purpose we adapted a baseline topic segmenter based on an HMM of topic states, each associated with a unigram LM that models topic-specific word distributions [17]. As in the previous tagging tasks, we extracted prosodic features around each potential boundary location, and let a decision tree compute posterior probabilities of the events (in this case, topic changes). By resampling the training events to a uniform distribution, we ensured that the posteriors are proportional to event likelihoods, as required for HMM integration [9, 15].

The results on this task are summarized in Table 3. We obtained a large, 24-27% relative error reduction from combining lexical and prosodic models. Also, similar to BN sentence segmentation, the prosodic model alone outperformed the LM. The prosodic features selected for topic segmentation were similar to those for sentence segmentation, but with more pronounced tendencies. For example,

TABLE 3
Topic segmentation weighted error on Broadcast News data. The evaluation metric used is a weighted combination of false alarm and miss errors [3].

Model	True words	Recognized words
LM only	0.1895	0.1897
Prosody only	0.1657	0.1731
Combined	0.1377	0.1438
Chance	0.300	0.300

TABLE 4
Dialog act classification error on highly ambiguous DA pairs in the Switchboard corpus.

Classification task Knowledge source	True words	Rec. words
Questions vs. Statements		
LM only	14.1	24.6
Prosody only	24.0	24.0
Combined	12.4	20.2
Agreements vs. Backchannels		
LM only	19.0	21.2
Prosody only	27.1	27.1
Combined	15.3	18.3
Chance	50.0	50.0

at the end of topic segments, a speaker tends to pause even longer and drop the pitch even closer to the baseline than at sentence boundaries.

3.3. Dialog act labeling in conversational speech. The third task we looked at was dialog act (DA) labeling. In this task the goal was to classify each utterance (rather than each word boundary) into a number of types, such as statement, question, acknowledgment, and backchannel. In [7] we investigated the use of prosodic features for DA modeling, alone and in conjunction with LMs. Prosodic features describing the whole utterance were fed to a decision tree. N-gram language models specific to each DA class provided additional likelihoods. These models can be applied to DAs in isolation, or combined with a statistical dialog grammar that models the contextual effects of nearby DAs. In a 42-way classification of Switchboard utterances, the prosody component improved the overall classification accuracy of such a combined model [11]. However, we found that prosodic features were most useful in disambiguating certain DAs that are particularly ambiguous based on their words alone. Table 4 shows results for two such binary DA discrimination tasks: distinguishing questions from statements, and backchannels ("uh-huh", "right") from agreements ("Right!"). Again, adding prosody boosted accuracy substantially over a word-only model. The features used for these and other DA disambiguation tasks, as might be expected, depend on the DAs involved, as described in [7].

3.4. Word recognition in conversational speech. All applications discussed so far had the goal of adding structural, semantic, or pragmatic information beyond what is contained in the raw word transcripts. Word recognition itself, however, is still far from perfect, raising the question: can prosodic cues be used to improve speech recognition accuracy? An early approach in this area was [16], using prosody to evaluate possible parses for recognized words, which in turn would be the basis for reranking word hypotheses. Recently, there have been a number of approaches that essentially condition the language model on prosodic evidence, thereby constraining recognition. The dialog act classification task mentioned above can serve this purpose, since many DA types are characterized by specific word patterns. If we can use prosodic cues to predict the DA of an utterance, we can then use a DA-specific LM to constrain recognition. This approach has yielded improved recognition in task-oriented dialogs [14], but significant improvements in large-vocabulary recognition remain elusive [11].

We have had some success using the hidden event N-gram model (previously introduced for sentence segmentation and disfluency detection) for word recognition [13]. As before, we computed prosodic likelihoods for each event type at each word boundary, and conditioned the word portion of the N-gram on those events. The result was a small, but significant 2% relative reduction in Switchboard word recognition error. This improvement was surprising given that the prosodic model had not been optimized for word recognition. We expect that more sophisticated and more tightly integrated prosodic models will ultimately make substantive contributions to word recognition accuracy.

3.5. Other corpora and tasks. We have recently started applying the framework described here to new types of data, including multiparty face-to-face meetings. We have found that speech in multiparty meetings seems to have properties more similar to Switchboard than to Broadcast News, with respect to automatic detection of target events [8]. Such data also offers an opportunity to apply prosody to tasks that have not been widely studied in a computational framework. One nice example is the modeling of turn-taking in meetings. In a first venture into this area, we have found that prosody correlates with the location and form of overlapping speech [8].

We also studied disfluency detection and sentence segmentation in the meeting domain, and obtained results that are qualitatively similar to those reported earlier on the Switchboard corpus [1]. A noteworthy result was that event detection accuracy on recognized words improved slightly when the models were trained on recognized rather than true words. This indicates that there is systematicity to recognition errors that can be partially captured in event models.

4. Conclusions. We have briefly summarized a framework for computational prosody modeling for a variety of tasks. The approach is based on modeling of directly measurable prosodic features and combination with lexical (statistical language) models. Results show that prosodic information can significantly enhance accuracy on several classification and tagging tasks, including sentence segmentation, disfluency detection, topic segmentation, dialog act tagging, and

overlap modeling. Finally, results so far show that speech recognition accuracy can also benefit from prosody, by constraining word hypotheses through a combined prosody/language model.

More information about individual research projects is available at http://www.speech.sri.com/projects/hidden-events.html, http://www.speech.sri.com/projects/sieve/, and http://www.clsp.jhu.edu/ws97/-discourse/.

REFERENCES

[1] D. BARON, E. SHRIBERG, AND A. STOLCKE, *Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues*, in Proceedings of the International Conference on Spoken Language Processing, J. H. L. Hansen and B. Pellom, eds., vol. 2, Denver, Sept. 2002, pp. 949–952.

[2] A. BATLINER, B. MÖBIUS, G. MÖHLER, A. SCHWEITZER, AND E. NÖTH, *Prosodic models, automatic speech understanding, and speech synthesis: toward the common ground*, in Proceedings of the 7th European Conference on Speech Communication and Technology, P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, eds., vol. 4, Aalborg, Denmark, Sept. 2001, pp. 2285–2288.

[3] G. DODDINGTON, *The Topic Detection and Tracking Phase 2 (TDT2) evaluation plan*, in Proceedings DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, Feb. 1998, Morgan Kaufmann, pp. 223–229. Revised version available from http://www.nist.gov/speech/tests/tdt/tdt98/.

[4] P. HEEMAN AND J. ALLEN, *Intonational boundaries, speech repairs, and discourse markers: Modeling spoken dialog*, in Proceedings of the 35th Annual Meeting and 8th Conference of the European Chapter, Madrid, July 1997, Association for Computational Linguistics.

[5] J. HIRSCHBERG AND C. NAKATANI, *Acoustic indicators of topic segmentation*, in Proceedings of the International Conference on Spoken Language Processing, R. H. Mannell and J. Robert-Ribes, eds., Sydney, Dec. 1998, Australian Speech Science and Technology Association, pp. 976–979.

[6] M. MAST, R. KOMPE, S. HARBECK, A. KIESSLING, H. NIEMANN, E. NÖTH, E. G. SCHUKAT-TALAMAZZINI, AND V. WARNKE, *Dialog act classification with the help of prosody*, in Proceedings of the International Conference on Spoken Language Processing, H. T. Bunnell and W. Idsardi, eds., vol. 3, Philadelphia, Oct. 1996, pp. 1732–1735.

[7] E. SHRIBERG, R. BATES, A. STOLCKE, P. TAYLOR, D. JURAFSKY, K. RIES, N. COCCARO, R. MARTIN, M. METEER, AND C. VAN ESS-DYKEMA, *Can prosody aid the automatic classification of dialog acts in conversational speech?*, Language and Speech, 41 (1998), pp. 439–487.

[8] E. SHRIBERG, A. STOLCKE, AND D. BARON, *Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation, disfluencies, and overlapping speech*, in Proceedings ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, M. Bacchiani, J. Hirschberg, D. Litman, and M. Ostendorf, eds., Red Bank, NJ, Oct. 2001, pp. 139–146.

[9] E. SHRIBERG, A. STOLCKE, D. HAKKANI-TÜR, AND G. TÜR, *Prosody-based automatic segmentation of speech into sentences and topics*, Speech Communication, 32 (2000), pp. 127–154. Special Issue on Accessing Information in Spoken Audio.

[10] K. SÖNMEZ, E. SHRIBERG, L. HECK, AND M. WEINTRAUB, *Modeling dynamic prosodic variation for speaker verification*, in Proceedings of the International Conference on Spoken Language Processing, R. H. Mannell and J. Robert-Ribes, eds., vol. 7, Sydney, Dec. 1998, Australian Speech Science and Technology Association, pp. 3189–3192.

[11] A. STOLCKE, K. RIES, N. COCCARO, E. SHRIBERG, D. JURAFSKY, P. TAYLOR, R. MARTIN, C. VAN ESS-DYKEMA, AND M. METEER, *Dialogue act modeling for automatic tagging and recognition of conversational speech*, Computational Linguistics, 26 (2000), pp. 339–373.

[12] A. STOLCKE, E. SHRIBERG, R. BATES, M. OSTENDORF, D. HAKKANI, M. PLAUCHÉ, G. TÜR, AND Y. LU, *Automatic detection of sentence boundaries and disfluencies based on recognized words*, in Proceedings of the International Conference on Spoken Language Processing, R. H. Mannell and J. Robert-Ribes, eds., vol. 5, Sydney, Dec. 1998, Australian Speech Science and Technology Association, pp. 2247–2250.

[13] A. STOLCKE, E. SHRIBERG, D. HAKKANI-TÜR, AND G. TÜR, *Modeling the prosody of hidden events for improved word recognition*, in Proceedings of the 6th European Conference on Speech Communication and Technology, vol. 1, Budapest, Sept. 1999, pp. 307–310.

[14] P. TAYLOR, S. KING, S. ISARD, AND H. WRIGHT, *Intonation and dialog context as constraints for speech recognition*, Language and Speech, 41 (1998), pp. 489–508.

[15] G. TÜR, D. HAKKANI-TÜR, A. STOLCKE, AND E. SHRIBERG, *Integrating prosodic and lexical cues for automatic topic segmentation*, Computational Linguistics, 27 (2001), pp. 31–57.

[16] N. M. VEILLEUX AND M. OSTENDORF, *Prosody/parse scoring and its applications in ATIS*, in Proceedings of the ARPA Workshop on Human Language Technology, Plainsboro, NJ, Mar. 1993, pp. 335–340.

[17] J. YAMRON, I. CARP, L. GILLICK, S. LOWE, AND P. VAN MULBREGT, *A hidden Markov model approach to text segmentation and event tracking*, in Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, vol. I, Seattle, WA, May 1998, pp. 333–336.