

COMBINING WORDS AND PROSODY FOR INFORMATION EXTRACTION FROM SPEECH

Dilek Hakkani-Tür

Gökhan Tür

Andreas Stolcke

Elizabeth Shriberg

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, U.S.A.
<http://www.speech.sri.com/>
{dilek,gokhan,stolcke,ees}@speech.sri.com

ABSTRACT

Information extraction from speech is a crucial step on the way from speech recognition to speech understanding. A preliminary step toward speech understanding is the detection of topic boundaries, sentence boundaries, and proper names in speech recognizer output. This is important since speech recognizer output lacks the usual textual cues to these entities (such as headers, paragraphs, sentence punctuation, and capitalization). Numerous *word*-based approaches to these tasks have been developed in the past; in this work we demonstrate the use of *prosodic* cues, alone and in combination with words, for segmentation and name finding. In experiments on the Broadcast News corpus, we find that prosodic cues alone allow sentence and topic segmentation that is at least as good as word-based methods alone, and that combining both types of cues gives significant wins. Named entity recognition, on the other hand, currently does not seem to benefit from prosodic cues, for several interesting reasons.

1. INTRODUCTION

Extracting particular types of entities, relations, or events from speech is a crucial step on the way from speech recognition to speech understanding. A preliminary step toward speech understanding is segmenting the recognizer output into sentences (sentence segmentation), stories (topic segmentation), and identifying proper names (named entity extraction).

Past automatic information extraction systems have depended mostly on lexical information [7, 1, 4, 6, 19, among others]. One problem for applying the text-based approach to speech input is the lack of typographic cues (such as headers, paragraphs, sentence punctuation, and capitalization). On the other hand, speech provides an additional, nonlexical knowledge source through its durational, intonational, and energy characteristics, i.e., its *prosody*.

Prosodic cues are known to be relevant to discourse structure in spontaneous speech [9, 8, 17, among others], and can therefore be expected to play a role in various information extraction tasks. Furthermore, prosodic cues by their nature are relatively unaffected by word identity, and should therefore improve the robustness of lexical information extraction methods based on automatic speech recognition.

Past studies involving prosodic information have generally relied on hand-coded cues (with the notable exception of Hirschberg and Nakatani [5]). We therefore believe the present work to be the first that combines fully automatic extraction of both lexical and prosodic information for information extraction. The general framework for combining lexical and prosodic cues for tagging speech with various kinds of *hidden* structural information is a further development of our earlier work on detecting sentence boundaries and disfluencies in spontaneous speech [16, 14, 12, 15].

2. PROSODIC MODELING

2.1. Data

For all tasks, the prosodic model used a wide range of features that were automatically extracted from about 70 hours (700 thou-

sand words) of the Linguistic Data Consortium (LDC) 1997 Broadcast News (BN) corpus. Sentence boundaries were automatically determined using the MITRE sentence tagger [10] based on capitalization and punctuation in the transcripts. Topic boundaries were derived from the SGML markup of the transcripts. named entity (NE) tags were obtained from hand-labeling efforts at MITRE and BBN. The word transcripts were aligned automatically with the acoustic waveforms to obtain pause and duration information. Raw pitch tracks were obtained using ESPS software. However, all our F0 features were based on post-processed pitch tracks based on an explicit model of pitch-halving/doubling, using straight-line stylizations for improved robustness [13].

2.2. Decision Trees

Let F_i be the features extracted from a window around the i th potential boundary or word, and let T_i be the boundary type (sentence/no-sentence or topic/no-topic) or NE type (person/location/organization/none) at that position. We trained CART-style decision trees [2] to predict the i th boundary or NE type, i.e., to estimate $P(T_i|F_i, W)$. The decision is only weakly conditioned on the word sequence W , insofar as some of the prosodic features depend on the phonetic alignment of the word models. We can thus expect the prosodic model estimates to be robust to recognition errors.

2.3. Feature Selection

We started with a large collection of features capturing the two major aspects of speech prosody, as in [12]:

- Duration: of pauses, final vowel and final rhymes, normalized both for phone durations and speaker statistics
- Pitch: F0 patterns, preceding the boundary, across the boundary, and pitch range relative to the speaker's baseline

We included features that, based on the descriptive literature, we believed should help the corresponding task. We developed versions of such features that could be defined at each interword boundary, and which could be extracted by completely automatic means (no human labeling). Furthermore, the features were designed to the extent possible to be independent of word identities, for use with recognizer output. The decision tree paradigm also allows us to add, and automatically select, other (nonprosodic) features that might be relevant to the task.

The greedy nature of the decision tree learning algorithm implies that larger initial feature sets can give worse results than smaller subsets. Furthermore, it is desirable to remove redundant features for computational efficiency and to simplify interpretation of results. For this purpose we developed an iterative feature selection algorithm to find useful task-specific feature subsets. The algorithm combined elements of brute-force search with previously determined heuristics about good groupings of features. We used the entropy reduction of the overall tree after cross-validation, as a method for selecting a good set of features. Entropy reduction is the difference in test-set entropy between the prior class distribution and the posterior distribution estimated by the tree; it is a more fine-grained metric than classification accuracy, and is also more relevant to the model

combination approach described later. The algorithm proceeds in two phases: in the first phase, the number of features is reduced, checking the effect of each feature on the performance by leaving out one feature at a time. The second phase then starts with the reduced number of features, and performs a beam search over all possible subsets of features.

3. HIDDEN MARKOV MODELING AND MODEL COMBINATION

We treated all three tasks as classification problems. The approach aims to find the classification T with highest probability given the information in W and F

$$\operatorname{argmax}_T P(T|W, F)$$

using statistical modeling techniques. Here F denotes the stream of prosodic features, T denotes the string of classifications,¹ and W denotes the string of words.

For all three tasks we employed a task-specific hidden Markov model (HMM) to represent the relation between the hidden variable T and the words, described in detail below. The models were trained in supervised fashion, i.e., using the labeled training data. As usual in HMM taggers, classification is achieved by finding the most likely HMM states given the observables (words and prosody). To incorporate the prosodic information into the HMM we modeled prosodic features as emissions from relevant HMM states, with likelihoods $P(F_i|T_i)$. We thus had to convert decision tree posterior probabilities $P(T_i|F_i)$ to likelihoods. This conversion can be accomplished either by dividing posteriors by prior probabilities, or by training the decision trees on a balanced training set. We preferred the resampling method, so the following relations hold:

$$P(T_i|F_i) = \frac{P(F_i|T_i)P(T_i)}{P(F_i)} \propto P(F_i|T_i)P(T_i) \propto P(F_i|T_i)$$

Note that $P(F_i)$ is a constant for different T_i , and $P(T_i)$ is also a constant by virtue of the resampling.

Our approach to combining word and prosodic knowledge sources assumes that the prosodic features depend only on the hidden variable in question (boundary/no-boundary, NE tag), but not on the words. We discuss various approaches around this strong assumption in [15].

3.1. Sentence Segmentation

For sentence segmentation, we relied on a hidden event N-gram language model (LM) of the type used in [14]. The states of the HMM consist of the end-of-sentence status of each word (boundary or no-boundary), plus preceding words and possibly boundary tokens to fill up the N-gram context ($N = 4$ in our experiments). Transition probabilities are given by the N-gram probabilities. HMM observations consist of the current word for each state, as well as the prosodic features. The word emission likelihoods are 1 by construction, whereas the prosodic likelihoods are obtained from decision trees as described above.

3.2. Topic Segmentation

For topic segmentation, we first constructed 100 individual unigram topic language models, using the multipass k -means algorithm described in [19]. We used the pooled Topic Detection and Tracking (TDT) Pilot and TDT-2 training data. We removed stories with fewer than 300 and more than 3000 words, leaving 19,916 stories with an average length of 538 words without any stop words. Then, similar to the Dragon topic segmentation approach [19], we built an HMM in which the states are topic clusters, and the observations are sentences (or chopped units). The resulting HMM forms a complete graph, allowing transition between any two topic clusters. In addition to the basic HMM segmenter, we incorporated two states for modeling the initial

¹ topic vs. nontopic for topic segmentation task, sentence vs. nonsense for sentence segmentation task, and NE type for name tagging

and final sentences of a topic segment. We reasoned that this can capture formulaic speech patterns used by broadcast speakers. Likelihoods for the start and end models are obtained as the unigram language model probabilities of the topic-initial and final sentences, respectively, in the training data. Note that a single start and end state are shared for all topics. Also, traversal of the initial and final states is optional in the HMM topology. We observed a 5% relative reduction in segmentation error over the baseline HMM topology using initial and final states.

Topic segmentation in the paradigm used by us and others [1] proceeds in two phases. In the first phase, the input is divided into contiguous strings of words assumed to belong to one topic each. We refer to this step as “chopping”. For example, in textual input, the natural units for chopping are sentences (as can be inferred from punctuation and capitalization). We could have used automatically determined sentence boundaries for chopping, but in a cheating experiment we found that pauses (thresholded at 0.66 second) give even better segmentation performance than true sentence boundaries. (For simplicity, we will use “sentence” to refer to units of chopping in the further description, regardless of the criterion used.) In the second phase, the sentences are further grouped into contiguous stretches belonging to one topic, i.e., the sentence boundaries are classified into “topic boundaries” and “nontopic boundaries”.²

As for sentence segmentation, we modified the word-based HMM to use probabilities from the decision tree estimator as additional likelihood scores, with an empirically optimized weighting. To this end, we inserted a fictitious *boundary* token between adjacent sentences, and introduced two more “boundary” states into the HMM topology. Between sentences, the model must visit one of the boundary states, denoting either the presence or absence of a topic boundary.

3.3. Named Entity Extraction

For name tagging, we used a simplified version of BBN’s name finder [7]. The states of the model were word/tag combinations, where the tag indicated whether a word was part of a proper name, and of what type (person, place, or organization). Transition probabilities consisted of bigram probabilities $P(w_1, t_1 | w_2, t_2)$ over these combined tokens. As in the sentence segmentation HMM, the word output for each state is given with probability 1. The prosodic model yields likelihoods for name vs. non-name classes, which are incorporated into the HMM as for the other tasks, as additional state likelihoods. For simplicity, and because we did not expect prosody to help with these classes, we extracted time, date, and numeric expressions (as defined in the DARPA Named Entity Extraction task) using (non-probabilistic) regular expression rules.

4. RESULTS AND DISCUSSION

For testing we used about three hours of the LDC 1997 BN corpus. We conducted experiments using transcribed words for all tasks, and where indicated additionally using 1-best recognition output from a simplified version of the SRI Hub-4 recognizer [11] (with a word error rate of 29%).

4.1. Sentence Segmentation

Table 1 summarizes our results on transcribed words for the sentence segmentation task. The baseline (or “chance”) performance for this task is 6.1% error, obtained by labeling all locations as nonboundaries. We investigated the following individual models: a stand-alone prosodic decision tree trained without downsampling (“prosody only”), a words-only HMM (“LM only”) trained on the same amount of data (700 thousand words) as the prosodic model, and a word-only HMM incorporating an additional 130 million words of training data (from the 1996 DARPA Hub-4 evaluation). This last model was motivated by the observation that word N-gram models typically require

² We do not consider the problem of detecting recurring, discontinuous instances of the same topic, a task known as “topic tracking” in the TDT paradigm.

far more training data than was available in the available acoustic corpus. We also created variants of these language models that incorporated turn boundaries (speaker changes) as pseudo-words, a cue that we knew from past work would be helpful. Finally, we combined each of the word language models with an equal-priors decision tree to investigate the effect of knowledge source combination. No optimized weighting of likelihoods was used.

Model	Error (%)	
	Model only	With prosody
Prosody only (700K)	4.3	
LM (700K)	5.7	3.9
LM (130M)	4.9	3.4
LM w/turns (700K)	4.8	3.6
LM w/turns (130M)	4.0	3.2

Table 1. Sentence segmentation performance (chance error rate is 6.1%). The LMs are differentiated by the number of words of training data.

Results show that the prosodic model alone performs better than a word-based language model trained on the same amount of data (with or without turn boundaries in the LM). Furthermore, a sizeable error reduction is achieved by combining the prosodic features with the word-based HMM, regardless of the LM used. This is interesting especially for the LM containing turn boundaries: turn information is also queried by the decision tree, thus violating the independence assumption made in the modeling.

Prosodic features. The feature selection identified six useful features for this task, which can be grouped into the following four classes. We provide for each group the relative frequency with which associated features are queried in the decision tree; this gives an approximate indication of feature importance.

1. Pause duration (49.3%). The duration of the nonspeech interval occurring at the boundary.
2. F0 differences across the boundary (23.3%). Several features compare the F0 following the boundary to F0 before the boundary. The F0s are measured over the duration of the words adjacent to the boundary, or over a fixed-length window of 200 milliseconds. Values are either mean F0, or minimum/maximum F0, in the regions surrounding the boundary. The mean captures a range effect; the minimum and maximum values make the measure more sensitive to local variation, such as rising to accented syllables, and final pitch falls.
3. Speaker change (23.2%). Whether or not a speaker change occurred at the boundary.
4. Rhyme duration (4.1%). Length of the rhyme (nucleus and coda) of the last syllable preceding the boundary.

4.2. Topic Segmentation

Table 2 gives our results using correct and recognized words for each of the individual knowledge sources as well as the combined model. All results reflect the word-averaged, weighted error metric used in the TDT-2 evaluations [18]. The TDT-2 metric gives more weight to false alarms than to misses, and our system was optimized accordingly using a twofold jackknifing of the test set.

As shown, the error rate for the prosody model alone is lower than that for the language model, and combining both models gives a further improvement. Using the combined model, the error rate decreased by 23% relative to the language model, for the correct words, and by 19% for recognized words.

Model	Weighted Error (%)	
	Correct Words	Recognized Words
Prosody only	18.9	19.9
Words only	19.5	20.9
Combined	15.0	16.8

Table 2. Summary of error rates using individual and combined models with pause duration as a chopping criterion.

Prosodic features. For the topic segmentation task, the iterative algorithm selected 5 features out of 73 potential features helpful for this task. Upon inspection, the following characteristics are modeled by the tree:

1. F0 differences across the boundary (44.0%)
2. Pause duration (36.3%)³
3. Speaker change (15.5%)
4. Gender (4.2%). We found stylistic differences between males and females in the use of F0 at topic boundaries. This is true even after proper normalization, for example, equating the gender-specific nontopic boundary distributions. In addition, we noted that nontopic pauses (i.e., chopping boundaries) are more likely to occur in male speech, a phenomenon that could have several causes and that awaits further analysis.⁴

4.3. Named Entity Extraction

We also looked into the effectiveness of prosodic modeling for the extraction of named entities. While most current systems use only word information for this task, we hypothesized that prosodic information might aid the task for the following reason. If named entities are well correlated with words carrying salient information, we would expect them to be prosodically prominent (showing pitch accents and increased duration). Thus a prosodic “prominence detector” could aid the NE task, particularly in the case of recognized words, by increasing the probability of an NE in regions that are acoustically prominent (and decreasing it where prominence is lacking).

Note, however, that a proposed win rests on the assumption of significant overlap between prominent words and named entities. However, inspection by a phonetics graduate student and later analyses revealed that there is only partial overlap between prominence and named entities. For example, named entities are often deaccented on repeated mention, as in “President Clinton ... Mr. Clinton ...” Conversely, content words carrying important information (such as “bomb” or “earthquake”) are frequent in news speech. These words are often prosodically prominent, but are not named entities. Still, since some win from prominence can be expected if the overlap between prominence and named entities is large enough, we built a prosodic classifier for this task and examined results for the tree alone as well as various combined models.

We did not expect prosodic modeling to be able to distinguish among different NE types since prominence should be similar across cases. Using prosody alone in an equal-priors tree for a binary classification task (NE of any type, versus non-NE; chance=50%) we found a moderate but significant win from prosody (accuracy=69%). The relative entropy reduction was much lower than that for the previous two tasks. Important features were, as expected, different from features helpful for sen-

³The importance of pause duration for this task is actually underestimated by this measure of feature use; as explained earlier, pause durations are already used during the chopping process, so that the decision tree is applied only to boundaries exceeding a certain duration. Separate experiments using boundaries below our chopping threshold show that the tree makes use of shorter pauses for segmentation decisions as well.

⁴For example, it could be that male speakers in BN are assigned longer topic segments on average, or that male speaker are more prone to pausing in general, or that males dominate the spontaneous speech portions where pausing is naturally more frequent.

tence and topic segmentation tasks. Whereas these segmentation tasks relied largely on pause and pitch differences across a boundary, prominence detection for named entities used mainly the durations of syllable nuclei and rhymes, and pitch differences within a word (to capture F0 excursions).

Model	Accuracy (%)			
	Content	Extent	Type	F-measure
Words only	91.81	87.06	90.91	89.93
Prosody only	23.62	13.95	13.83	17.13
Combined	91.92	87.45	91.16	90.18

Table 3. Name tagging performance. *Type* checks for the correctness of NE type for each word; *content* evaluates the performance on the correctly recognized words, *extent* checks for the boundaries of the named entity, and *F-measure* is a combination of these three measures [3].

However, the small win for the tree alone was of no help in named entity detection above and beyond that of the language model, as shown in Table 3. Alternative methods of training the prosodic classifier, such as using first mentions (via an imperfect heuristic based on exact repetition of a name) or of varying the scoring for the span of a name, did not yield better results.

In a further, diagnostic, experiment we found that if we removed function words from our tree-only analysis, the small win we observed earlier (69% on an equal-priors tree) disappeared. This suggests that the better-than-chance result from prosody came from distinguishing content words from function words, rather than named-entity content words from non-named-entity content words. Thus it suggests that prominence patterns may not differ for named entities and other “important” content words. Although results require further inspection, we hypothesize that given the current definition of named entities, it may be difficult to find a win from prosody because of a lack of good overlap between the words that speakers deem important, and those that are defined as named entities for this task.

5. CONCLUSIONS

We have developed a combined approach to several tasks, relevant to speech understanding and information extraction from speech, that enhance standard word-based taggers and classifiers with prosodic information. We investigated the sentence segmentation, topic segmentation, and name-finding tasks in the Broadcast News domain. Results on the two segmentation tasks show excellent performance for prosody-only models (competitive with word-only models), and significantly improved performance (20-25% error reduction) when combining prosodic and word-based models. No such win was observed for the name-finding task, which is most likely due to insufficient overlap between “names” and prosodically prominent “important” words.

ACKNOWLEDGMENTS

We thank Becky Bates, Ze’ev Rivlin, Ananth Sankar, and Kemal Sönmez for invaluable assistance in preparing the data for this study, as well as for many helpful discussions. This research was supported by DARPA and NSF under NSF grant IRI-9619921 and DARPA contract no. N66001-97-C-8544. The views herein are those of the authors and should not be interpreted as representing the policies of the funding agencies.

REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, Lansdowne, VA, 1998.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA, 1984.

[3] J. D. Burger, D. Palmer, and L. Hirschman. Named entity scoring for speech input. In *Proc. COLING-ACL*, vol. 1, pp. 201–205, Montreal, 1998.

[4] M. A. Hearst. TextTiling: Segmenting text into multiparagraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[5] J. Hirschberg and C. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. ACL*, pp. 286–293, Santa Cruz, CA, 1996.

[6] H. Kozima. Text segmentation based on similarity between words. In *Proc. ACL*, pp. 286–288, Ohio State University, Columbus, Ohio, 1993.

[7] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel. Named entity extraction from speech. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, 1998.

[8] D. J. Litman and R. J. Passonneau. Combining multiple knowledge sources for discourse segmentation. In *Proc. ACL*, pp. 108–115, MIT, Cambridge, MA, 1995.

[9] S. Nakajima and J. F. Allen. A study on prosody and discourse structure in cooperative dialogues. *Phonetica*, 50:197–210, 1993.

[10] D. D. Palmer and M. A. Hearst. Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics*, 23(2):241–267, 1997.

[11] A. Sankar, F. Weng, Z. Rivlin, A. Stolcke, and R. R. Gadde. The development of SRI’s 1997 Broadcast News transcription system. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 91–96, Lansdowne, VA, 1998.

[12] E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proc. EUROSPEECH*, vol. 5, pp. 2383–2386, Rhodes, Greece, 1997.

[13] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, vol. 7, pp. 3189–3192, Sydney, 1998. Australian Speech Science and Technology Association.

[14] A. Stolcke and E. Shriberg. Automatic linguistic segmentation of conversational speech. In H. T. Bunnell and W. Idsardi, editors, *Proc. ICSLP*, vol. 2, pp. 1005–1008, Philadelphia, 1996.

[15] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, vol. 5, pp. 2247–2250, Sydney, 1998. Australian Speech Science and Technology Association.

[16] A. Stolcke, E. Shriberg, D. Hakkani-Tür, G. Tür, Z. Rivlin, and K. Sönmez. Combining words and speech prosody for automatic topic segmentation. In *Proceedings DARPA Broadcast News Workshop*, Herndon, VA, 1999.

[17] M. Swerts and M. Ostendorf. Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication*, 22(1):25–41, 1997.

[18] The Topic Detection and Tracking Phase 2 (TDT2) evaluation plan. Available from <http://www.nist.gov/speech/tdt98/tdt98.htm>, 1998.

[19] J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A hidden Markov model approach to text segmentation and event tracking. In *Proc. ICASSP*, vol. 1, pp. 333–336, Seattle, WA, 1998.