ACOUSTIC INDICATORS OF TOPIC SEGMENTATION

Julia Hirschberg *Christine Nakatani*

AT&T Labs — Research
Florham Park, NJ
USA

ABSTRACT

The segmentation of text and speech into topics and subtopics is an important step in document interpretation. For text, formatting information, such as headings and paragraphing, is available to aid in this endeavor, although this information is by no means sufficient. For speech, the task is even more difficult. We present results of the application of machine learning techniques to the automatic identification of intonational phrases beginning and ending 'topics' determined independently by annotators for two corpora — the Boston Directions Corpus and the Broadcast News (HUB-4) DARPA/NIST database.

1. INTRODUCTION

The segmentation of speech into meaningful units of analysis is an important step in the interpretation of the intonational features speakers use, as well as of other linguistic features derived from text analysis. It has been hypothesized that the segmentation of speech is useful to spoken language interpretation tasks, such as automatic speech recognition and discourse analysis, because there exist fundamental correspondences between the acoustic-prosodic and other linguistic structures of spoken language.

Most theoretical models of intonation assume one or more levels of intonation phrasing, under which tonal feature variation is interpreted. Intuitively, the intonational phrasing of an utterance divides it into meaningful 'chunks' of information. For example, variation in a sentence's intonational phrasing can change the meaning that hearers are likely to assign to an individual utterance of the sentence. The sentence "*Bill doesn't drink because he's unhappy*", for example, can be produced in two ways to convey distinct interpretations, depending upon whether or not a phrase boundary occurs between *drink* and *because*: without a boundary, Bill drinks, but not because he is unhappy; with a boundary, Bill doesn't drink at all. At a lower level of linguistic structure, that of morphological analysis, it is generally assumed that intonation phrasing does not occur in the middle of a single morphological unit. This property makes the intonational phrase a suitable unit for ASR analysis, since it is likely that a single word will not cross over an intonational phrase boundary. At a higher level of linguistic structure, namely that of discourse structure analysis, it has been demonstrated that acoustic-prosodic properties of intonational phrases are correlated with their discourse structural position, e.g. [8, 3].

In this paper, we present results on the identification of intonational phrase boundaries from a small set of acoustic features, using a machine learning technique, Classification and Regression Trees (CART) [2], that builds decision-trees from vectors of independent variables, each associated with a dependent variable. Our training and testing corpora are the Boston Directions Corpus of task-oriented monologue speech and the HUB-IV Broadcast News database of monologue and multi-party news speech. Our goals are to provide intonational phrase segmentation as a front end for an ASR engine and to infer topic structure from acoustic-prosodic features. These efforts are aimed at improving the ease and flexibility of retrieving and browsing speech documents from a large audio database.

2. MACHINE LEARNING EXPERIMENTS

The current segmentation procedures were trained and tested on a corpus of read and spontaneous speech, the Boston Directions Corpus.[1] The subcorpus used for the segmentation experiments comprises elicited monologues produced by four non-professional speakers, three male and one female, who were given written instructions to perform a series of nine increasingly complex direction-giving tasks. Speakers first explained simple routes such as getting from one station to another on the subway, and progressed gradually to the most complex task of planning a round-trip journey from Harvard Square to several Boston tourist sights. Thus, the tasks were designed to require increasing levels of planning complexity. The spontaneous, elicited speech was subsequently orthographically transcribed, with false starts and other speech errors repaired or omitted; subjects returned several weeks after their first recording to read aloud from transcriptions and these read productions were recorded and analyzed as well. For earlier studies, a prosodic transcription of the speech had also been made by hand, using the ToBI standard for prosodic transcription [6]. This transcription provides us with a breakdown of the speech sample into intonational phrases. There were a total of 3306 intonational phrases, 1292 in the read speech and 2014 in the spontaneous. A second, blind test corpus was a portion of the ARPA/NIST HUB-IV or Broadcast News database that is being used for the TREC information retrieval from speech task. This corpus consists of recorded news programs, containing multi-speaker professional and non-professional read and spontaneous speech. Speech from this corpus was hand-labeled for testing.

[1] This corpus was designed and collected by the authors in collaboration with Barbara Grosz at Harvard University.

For this study, several acoustic-prosodic measurements were used as predictive features for identifying whether a short 10 msec frame of the speech signal occurred within an intonational phrase (INPHRASE) or in the break between two intonational phrases (INBREAK). The length of the frame was chosen based on the method used for calculating measurements, namely automatic pitch-tracking. All recordings in the development and test corpora were down-sampled to 16K, then filtered at 70 Hz with DC offsets removed, and pitch-tracked using a 10 msec frame using the Entropic pitch-tracker, *getf0*. These non-overlapping frames provided the unit for the dependent variable in our segmentation experiments, i.e. the binary classification of each frame as INPHRASE or INBREAK.

Independent variables were selected from the range of possible outputs of the pitch tracker, *getf0*, which provides four types of information per frame: an estimate of the fundamental frequency (f0), a binary flag denoting the program's estimation of the probability of voicing (pvoice), root mean squared energy (rms),[2] and ac-peak, the peak normalized cross-correlation value found by the program to determine the f0 estimate.[3] We also conducted experiments with output from an earlier Entropic pitch-tracker, *formant*, which provides continuous, though less accurate, estimations of the probability of voicing. These variables were available to the machine learning algorithms in absolute and normalized forms (normalized by mean, maximum, and minimum values for the speech under analysis), and also in ratios of the value of the prior frame to that of the current frame. We tested measurements taken over a variety of different-sized frame windows, from 1 to 27 frames in length, to make contextual information available to the algorithms. Experiments were run by partitioning our training corpus by individual speaker and by speaking style (i.e. read versus spontaneous productions), to identify models that best predicted new data. We then tested our speaker/style models on two tasks: (1) recall of discourse segment boundaries on the Boston Directions Corpus and (2) phrase prediction on the HUB-IV data.

3. MODEL DEVELOPMENT

Models that classify frames with 87-93% accuracy were developed from the training corpus in several stages. First, we identified the best performing acoustic feature sets predicting the current frame based only on itself and at most a single frame of context. This contextual information included the absolute acoustic information (e.g. f0, rms) for these contextual frames, or the relative difference between such acoustic values for the previous the current frame. Next, models for this best feature set of SINGLE-FRAME-BASED features were trained on each speaker and each speaking style in our corpus, and tested on all other partitions of speaker and style. This cross-speaker, cross-style testing procedure revealed speaker/style models that best modeled the other speaker/style data in our training corpus.[4] In the second stage, the training data partition for the speaker/style model that best predicted the other speaker/style data was used to select a distinct feature set of contextual, or MULTI-FRAME-BASED, features. Contextual features were computed over windows varying from 2 to 27 frames in length. In addition, these windows were aligned with the current frame being classified in three different ways. Specifically, we examined windows that were centered on, or aligned with the left or right edge of the frame to be classified. The best-performing window size for each feature in each alignment was selected at this stage. The models developed at this stage of testing in general were computationally considerably more expensive than the single-frame-based feature models. In the third stage, the best combination of single-frame-based and multi-frame window-based features was identified. This was determined by comparing predictions of models created by all possible combinations of single-frame-based and multi-frame-based features that were identified as good predictors in the first two stages.

Finally, this combined model of frame-based and multi-frame-window-based features was tested in two experiments. The first was aimed at the goal of inferring topic structure from acoustic-prosodic features. The second was aimed at assessing the potential value of using predicted phrase boundaries for audio browsing applications, to help the user navigate through the audio as well as to enable the playback of prosodically well-formed speech units. We note that one of the models we trained is currently being used to segment a large speech corpus of broadcast news programs into manageable segments as a front end for an ASR engine under development at AT&T Labs – Research. While the evaluation of this usage of our tool, e.g. its possible effect on recognition accuracy rates, remains to be done, this phrase segmenter improves the computational efficiency over methods using fixed overlapping intervals of speech.

Sets of single-frame-based features were studied systematically by performing cross-validated training on each of eight partitions of the BDC corpus (4 x individual speaker and 2 x speaking style), using CART. The best-performing feature set for this stage of development included: normalized mean f0, f0 ratio, pvoice for previous frame, pvoice, pvoice for subsequent frame, normalized mean rms, rms ratio and ac-peak. Using this single-frame-based model, exhaustive cross-speaker, cross-style testing was carried out among the partitions by training on all data from one partition and testing on a subportion of each of the remaining partitions. This testing procedure revealed speaker/style models that best modeled each other. A second analysis of the cross-speaker, cross-style results was made, based on precision and recall measures for each of the binary classifications. These results led to the selection of the partition on which to develop the multi-frame-based feature models. Due to the nature of our target application, audio browsing, we opted to maximize precision of the INBREAK class, since for both ASR and control of speech playback, it would be preferrable to err on the side of providing longer phrases instead of risking the placement of infelicitous phrase breaks in, say, mid-word. Error analysis indeed revealed that the majority of false positives occurred in very short equences of frames classified as INBREAK. However, with too low recall, intonational phrases would be very long, which might also compromise the usefulness of intonational segmentation for these same

[2] The rms value for each frame is computed based on a 30 msec hanning window with its left edge 5 msec before the beginning of the frame.

[3] In unvoiced regions, this is the largest cross-correlation value found at any lag.

[4] Prediction models trained on combinations of various speaker/style partitions did not prove superior to models trained on a single speaker/style data partition. Therefore, we report results for models trained on single speaker/style data partitions.

applications. The chosen model demonstrated average recall of .63, with precision of .79.

This model was employed in the testing of three types of contextual features for each acoustic variable (e.g. f0, rms). Contexts were defined as windows, 2 to 27 frames in length, that were (1) centered on (2) aligned with the left edge (left context features), or (3) aligned with the right edge (right context features) of the frame being classified. Multi-frame-based features were computed for all contexts for three of the best-performing acoustic measures discovered in prior testing: normalized mean f0, normalized mean rms, and probability of voicing (pvoice). The strategy we used was to train on a single acoustic feature in a single window position at a time, incrementing the window size until maximum performance was realized on test data from other speakers. This determined the best window size for each feature in each of the three alignment positions.

These multi-frame-based features, computed over their optimal window sizes, were then combined in all possible ways with the best-predicting single-frame-based features, to establish the final, best-performing model combining contextual and non-contextual features. The predictors occurring in the final COMBINED MODEL included two multi-frame-based features and one single-frame-based feature: 15 frame centered window of normalized mean rms, 19 frame left context window of normalized mean f0, and ac-peak for the current frame alone. Performance figures on new data for the combined model built using CART ranged from .80 cross-validated estimate for our worst test set to .93 for our best. Adding multi-frame-based features representing acoustic information about the current frame's context improved classification accuracy by 2-5% per test partition over the single-frame-based models. This new model is also more concise than the earlier best-predicting models, making use of only three acoustic predictors. While energy is represented by a window-based feature centered on the frame and f0 is represented by a left-context window-based feature in the combined model, no window-based feature calculated for the right context was found to be useful in a combined model.

4. TESTING

The feature set used in the combined model of single-frame-based and contextual features described in the previous section was applied in two experiments. The first experiment was aimed at inferring discourse segment structure from acoustic-prosodic features. The second was aimed at assessing the potential value of using predicted phrase boundaries for audio browsing applications, to help the user navigate through the audio as well as to enable the playback of prosodically well-formed speech units.

Although intonational phrase labels were used to derive training examples for the above systems, the systems classified each 10 msec frame of speech as belonging to a phrase or to a break between phrases, and did not directly identify phrase boundaries as such. We needed first to determine if, from such sequences of frame classifications, it would be possible to derive higher-level segmentations, such as intonational phrase boundaries and discourse segment boundaries. Evaluation on separate training and testing partitions of the Boston Directions Corpus provides preliminary answers to this question.

The feature set from the combined frame classification model was trained on one speaker to produce the frame-based classifications described above. To infer segmental structure beyond the frame level, we utilized a basic smoothing procedure which simply filtered out sequences of frame breaks by two means: setting a minimum sequence length, say 3 frames, or preserving a set ratio, say 20%, of all identified frame break sequences with longer sequences preferred. Then, higher-level segments were inferred by taking all filtered sequences of frame breaks to represent segment boundary markers.

First, we evaluated this procedure on reliable (consensus) discourse segment boundary labels obtained by three labelers in an earlier study.[5] Table 1 illustrates performance and recall figures for discourse segment boundaries, at varying filter ratios. The ratios were selected based on analysis of a separate set of discourse segment data (from naive labelers on h3s using the methods in [4]), in which 24% of intonational phrases were marked as consensus discourse segment beginnings. The h1s test set contained 85 true boundaries. As can be seen, there are significant

filter ratio	classified INBRK	true hits	recall	precision
.08	85	35	.412	.412
.16	171	50	.588	.292
.24	253	55	.647	.217
.32	340	59	.694	.174

Table 1: Evaluation of combined model on discourse segment boundary identification on the Boston Directions Corpus.

trade-offs between recall and precision. It appears that it will be important to optimize these trade-offs depending on the target application or on the performance of complementary classification systems, such as text-based models, that have different coverage of the data.

Using the same filter ratios, we evaluated the model's ability to identify intonational phrase boundaries in the h1s test set, of which there were 366 labeled boundaries. Results are given in Table 2. The results are almost the

filter ratio	classified INBRK	true hits	recall	precision
.08	85	67	.183	.788
.16	171	130	.355	.760
.24	253	176	.481	.696
.32	340	203	.555	.597

Table 2: Evaluation of combined model on intonational phrase boundary identification on the Boston Directions Corpus.

inverse of the discourse segment boundary figures, suggesting that trade-offs need to be considered together. For example, a low filter ratio (.08) will provide acceptable

[5] Segmenters were experts who listened to the speech while segmenting. The averaged kappa score for inter-labeler agreement was .80 for the data used. Further details on the methods used for discourse segment data collection are provided in [3].

segmentation for our purposes and will find a reasonable minority of discourse segment boundaries, but a higher ratio, say .16, will find nearly twice as many discourse segment boundaries with only a 3% decrease in precision for intonational phrases.

There is clearly room for improvement in the performance of our frame classifier and of our smoothing procedure, and we intend to experiment with data partitioning techniques to increase the size of a reliable training corpus. However, in our testing described above, we used a rather conservative measure to determine when a frame sequence matched a phrase boundary: if the midpoint of the true phrase break (computed from the ToBI labels) fell within a frame break sequence, the two boundaries were said to match. If we employ a looser criterion, allowing the existence of any overlapping frames to constitute a match, precision and recall figures improve by up to 6%. For audio browsing applications, such a metric may be quite acceptable.

We have begun similar evaluation of intonational phrase identification on the HUB-IV corpus. Using the combined feature set, we trained models for each speaker/style data partition in the Boston Directions Corpus. These models were tested on an initial test corpus consisting of 230 sec of a National Public Radio broadcast, containing professionally read speech from two speakers (one male, one female) made up of 88 intonational phrases (identified by hand-labeling). In this experiment, no independent estimate could be made of the expected frequency of intonational phrases on the test set, so we utilized a minimum break frame sequence length of 3 (i.e. 30 msec) to smooth the output of frame classifications. Performance figures are given in Table 3. The results reported in Table 3 are computed using the relaxed scoring method that counts two boundaries as matching if they share at least one overlapping frame. Amongst our eight partitions, the read speaking style models for h3 and h4 deliver strong recall results of .92 and .95 respectively, and reasonable precision (.74 and .71). The highest precision of .79, however, is achieved by the h2s model. Finally, as may have been expected from model development results on the Boston Directions Corpus, models trained on h1r and h4s partitions give substantially lower performance, as evidenced by the precision scores of .48 and .44 for these models.

Train set*	identified phrases	recall	precision
h1r	43	.49	.48
h1s	70	.80	.74
h2r	70	.80	.69
h2s	67	.76	.79
h3r	81	.92	.74
h3s	72	.82	.77
h4r	84	.95	.71
h4s	75	.85	.44

* h?=speaker ID, r=Read and s=Spontaneous

Table 3: Evaluation of the speaker/style combined models on intonational phrase boundary identification for broadcast news.

5. DISCUSSION

These initial experiments suggest that the identification of intonational phrasing by purely automatic means is feasible. Given the limited amount of training data utilized, the performance especially on the HUB-IV blind test set suggests it is a useful approach as a front end for an ASR engine and for audio browsing applications. To help us determine practical minimum threshholds of performance, we are experimentating with the use of our phrase identification system in several audio browsing interfaces. To improve performance further, we need to explore several technical issues in machine learning, such as (1) approaches to meta-learning, to automatically partition our training data in more sophisticated ways; and (2) the problem of learning unified representations of hierarchically structured concepts, such as our hypothesized hierarchy of frames, intonational phrases and discourse segments. Currently, our multi-layer classification approach presents us with many design choices that concern the integration of various machine learning systems and techniques into a multi-pass "architecture" of sorts for intonational phrase prediction. Formalizing and constraining such a design in a coherent machine learning framework for learning complex, hierarchically dependent structures incrementally is an important goal for our further research.

6. REFERENCES

1. D. Bolinger. *Intonation and Its Uses: Melody in Grammar and Discourse.* Edward Arnold, London, 1989.
2. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks, Pacific Grove CA, 1984.
3. J. Hirschberg and C. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting*, Santa Cruz, 1996. Association for Computational Linguistics.
4. C. Nakatani, B. Grosz, and H. Julia. Discourse structure in spoken language: Studies on speech corpora. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford, March 1995. AAAI.
5. J. B. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology, September 1980. Distributed by the Indiana University Linguistics Club.
6. J. Pitrelli, M. Beckman, and J. Hirschberg. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the Third International Conference on Spoken Language Processing*, volume 2, pages 123–126, Yokohama, 1994. ICSLP.
7. M. D. Riley. Some applications of tree-based modelling to speech and language. In *Proceedings of the Speech and Natural Language Workshop*, Cape Cod MA, October 1989. DARPA, Morgan Kaufmann.
8. M. Swerts, R. Gelyukens, and J. Terken. Prosodic correlates of discourse units in spontaneous speech. In *Proceedings*, pages 421–428, Banff, October 1992. International Conference on Spoken Language Processing.