Combining Prosodic and Text Features for Segmentation of Mandarin Broadcast News

Gina-Anne Levow
University of Chicago
levow@cs.uchicago.edu

Abstract

Automatic topic segmentation, separation of a discourse stream into its constituent stories or topics, is a necessary preprocessing step for applications such as information retrieval, anaphora resolution, and summarization. While significant progress has been made in this area for text sources and for English audio sources, little work has been done in automatic, acoustic feature-based segmentation of other languages. In this paper, we consider exploiting both prosodic and text-based features for topic segmentation of Mandarin Chinese. As a tone language, Mandarin presents special challenges for applicability of intonation-based techniques, since the pitch contour is also used to establish lexical identity. We demonstrate that intonational cues such as reduction in pitch and intensity at topic boundaries and increase in duration and pause still provide significant contrasts in Mandarin Chinese. We build a decision tree classifier that, based only on word and local context prosodic information without reference to term similarity, cue phrase, or sentence-level information, achieves boundary classification accuracy of 84.6-95.6% on a balanced test set. We contrast these results with classification using text-based features, exploiting both text similarity and n-gram cues, to achieve accuracies between 77-95.6%, if silence features are used. Finally we integrate prosody, text, and silence features using a voting strategy to combine decision tree classifiers for each feature subset individually and all subsets jointly. This voted decision tree classifier yields an overall classification accuracy of 96.85%, with 2.8% miss and 3.15% false alarm rates on a representative corpus sample, demonstrating synergistic combination of prosodic and text features for topic segmentation.

1 Introduction

Natural spoken discourse is composed of a sequence of utterances, not independently generated or randomly strung together, but rather organized according to basic structural principles. This structure in turn guides the interpretation of individual utterances and the discourse as a whole. Formal written discourse signals a hierarchical, tree-based discourse structure explicitly by the division of the text into chapters, sections, paragraphs, and sentences. This structure, in turn, identifies domains for interpretation; many systems for anaphora resolution rely on some notion of locality (Grosz and Sidner, 1986). Similarly, this structure represents topical organization, and thus would be useful in information retrieval to select documents where the primary sections are on-topic, and, for summarization, to select information covering the different aspects of the topic.

Unfortunately, spoken discourse does not include the orthographic conventions that signal structural organization in written discourse. Instead, one must infer the hierarchical structure of spoken discourse from other cues. Prior research (Nakatani et al., 1995; Swerts, 1997) has shown that human labelers can more sharply, consistently, and confidently identify discourse structure in a word-level transcription when an original audio recording is available than they can on the basis of the transcribed text alone. This finding indicates that substantial additional information about the structure of the discourse is encoded in the acoustic-prosodic features of the utterance. Given the often errorful transcriptions available for large speech corpora, we choose to focus here on fully exploiting the prosodic cues to discourse structure present in the original speech. We then compare the effectiveness of a pure prosodic classification to text-based and mixed text and prosodic based classification.

In the current set of experiments, we concentrate on sequential segmentation of news broadcasts into individual stories. While a richer hierarchical segmentation is ultimately desirable, sequential story segmentation provides a natural starting point. This level of segmentation can also be most reliably performed by human labelers and thus can be considered most robust, and segmented data sets are pub-

licly available.

Furthermore, we apply prosodic-based segmentation to Mandarin Chinese, in addition to textual features. Not only is the use of prosodic cues to topic segmentation much less well-studied in general than is the use of text cues, but the use of prosodic cues has been largely limited to English and other European languages.

2 Related Work

Most prior research on automatic topic segmentation has been applied to clean text only and thus used textual features. Text-based segmentation approaches have utilized term-based similarity measures computed across candidate segments (Hearst, 1994) and also discourse markers to identify discourse structure (Marcu, 2000).

The Topic Detection and Tracking (TDT) evaluations focused on segmentation of both text and speech sources. This framework introduced new challenges in dealing with errorful automatic transcriptions as well as new opportunities to exploit cues in the original speech. The most successful approach (Beeferman et al., 1999) produced automatic segmentations that yielded retrieval results approaching those with manual segmentations, using text and silence features. (Tur et al., 2001) applied both a prosody-only and a mixed text-prosody model to segmentation of TDT English broadcast news, with the best results combining text and prosodic features. (Hirschberg and Nakatani, 1998) also examined automatic topic segmentation based on prosodic cues, in the domain of English broadcast news, while (Hirschberg et al., 2001) applied similar cues to segmentation of voicemail.

Work in discourse analysis (Nakatani et al., 1995; Swerts, 1997) in both English and Dutch has identified features such as changes in pitch range, intensity, and speaking rate associated with segment boundaries and with boundaries of different strengths. They also demonstrated that access to acoustic cues improves the ease and quality of human labeling.

3 Prosody and Mandarin

In this paper we focus on topic segmentation in Mandarin Chinese broadcast news. Mandarin Chinese is a tone language in which lexical identity is determined by a pitch contour - or *tone* - associated with each syllable. This additional use of pitch raises the question of the cross-linguistic applicability of the prosodic cues, especially pitch cues, identified for non-tone languages. Specifically, do we find intonational cues in tone languages? The fact that emphasis is marked intonationally by expansion of pitch range even in the presence of Mandarin lexical tone (Shen, 1989) suggests encouragingly that prosodic, intonational cues to other aspects of information structure might also prove robust in tone languages.

4 Data Set

We utilize the Topic Detection and Tracking (TDT) 3 (Wayne, 2000) collection Mandarin Chinese broadcast news audio corpus as our data set. Story segmentation in Mandarin and English broadcast news and newswire text was one of the TDT tasks and also an enabling technology for other retrieval tasks. We use the segment boundaries provided with the corpus as our gold standard labeling. Our collection comprises 3014 news stories drawn from approximately 113 hours over three months (October-December 1998) of news broadcasts from the Voice of America (VOA) in Mandarin Chinese, with 800 regions of other program material including musical interludes and teasers. The transcriptions span approximately 750,000 words. Stories average approximately 250 words in length to span a full story. No subtopic segmentation is performed. The audio is stored in NIST Sphere format sampled at 16KHz with 16-bit linear encoding.

5 Prosodic Features

We consider four main classes of prosodic features for our analysis and classification: pitch, intensity, silence and duration. Pitch, as represented by f0 in Hertz was computed by the "To pitch" function of the Praat system (Boersma, 2001). We selected the highest ranked pitch candidate value in each voiced region. We then applied a 5-point median filter to smooth out local instabilities in the signal such as vocal fry or small regions of spurious doubling or halving. Analogously, we computed the intensity in decibels for each 10ms frame with the Praat "To intensity" function, followed by similar smoothing.

For consistency and to allow comparability, we compute all figures for word-based units, using the automatic speech recognition transcriptions provided with the TDT Mandarin data. The words are used to establish time spans for computing pitch or intensity mean or maximum values, to enable durational normalization and the pairwise comparisons reported below, and to identify silence or non-speech duration.

It is well-established (Ross and Ostendorf, 1996) that for robust analysis pitch and intensity should be normalized by speaker, since, for example, average pitch is largely incomparable for male and fe-

male speakers. In the absence of speaker identification software, we approximate speaker normalization with story-based normalization, computed as $\frac{val-mean}{mean}$, assuming one speaker per topic[1]. For duration, we consider both absolute and normalized word duration, where average word duration is used as the mean in the calculation above.

6 Prosodic Analysis

To evaluate the potential applicability of prosodic features to story segmentation in Mandarin Chinese, we performed some initial data analysis to determine if words in story-final position differed from the same words used throughout the story in news stories. This lexical match allows direct pairwise comparison. We anticipated that since words in Mandarin varied not only in phoneme sequence but also in tone sequence, a direct comparison might be particularly important to eliminate sources of variability. Features that differed significantly would form the basis of our classifier feature set.

We found significant differences for each of the features we considered. Specifically, word duration, normalized mean pitch, and normalized mean intensity all differed significantly for words in topic-final position relative to occurrences throughout the story (paired t-test, two-tailed, $p < 0.05, p < 0.0025, p < 0.0025$, respectively) . Word duration increased, while both pitch and intensity decreased. A small side experiment using 15 hours of English broadcast news from the TDT collection shows similar trends, though the magnitude of the change in intensity is smaller than that observed for the Chinese. Furthermore, comparison of average pitch and average intensity for 1, 5, and 10 word windows at the beginning and end of news stories finds that pitch and intensity are both significantly higher ($p < 0.001$) at the start of stories than at the end.

These contrasts are consistent with, though in some cases stronger than, those identified for English (Nakatani et al., 1995) and Dutch (Swerts, 1997). The relatively large size of the corpus enhances the salience of these effects. We find, importantly, that reduction in pitch as a signal of topic finality is robust across the typological contrast of tone and non-tone languages. These findings demonstrate highly significant intonational effects even in tone languages and suggest that prosodic cues may be robust across wide ranges of languages.

[1]This is an imperfect approximation as some stories include off-site interviews, but seems a reasonable choice in the absence of automatic speaker identification.

7 Classification

7.1 Prosodic Feature Set

The results above indicate that duration, pitch, and intensity should be useful for automatic prosody-based identification of topic boundaries. To facilitate cross-speaker comparisons, we use normalized representations of average pitch, average intensity, and word duration. We also include absolute word duration. These features form a word-level context-independent feature set.

Since segment boundaries and their cues exist to contrastively signal the separation between topics, we augment these features with local context-dependent measures. Specifically, we add features that measure the change between the current word and the next word.[2] This contextualization adds four contextual features: change in normalized average pitch, change in normalized average intensity, change in normalized word duration, and duration of following silence or non-speech region.

7.2 Text Feature Set

In addition to the prosodic features which are our primary interest, we also consider a set of features that exploit textual similarity to identify segment boundaries. Motivated by text and topic similarity measures in the vector space model of information retrieval (Salton, 1989), we compute a vector representation of the words in 50 word windows preceding and following the current potential boundary position. We compute the cosine similarity of these two vectors. We employ a $tf * idf$ weighting; each term is weighted by the product of its frequency in the window (tf) and its inverse document frequency (idf) as a measure of topicality. We also consider the same similarity measure computed across 30 word windows. The final text similarity measure we consider is simple word overlap, counting the number of words that appear in both 50 word windows defined above. We did not remove stopwords, and used the word-based units from the ASR transcription directly as our term units. We expect that these measures will be minimized at topic boundaries where changes in topic are accompanied by changes in topical terminology.

Finally we identified a small set of word unigram features occuring within a ten-word window immediately preceding or following a story boundary that

[2]We have posed the task of boundary detection as the task of finding segment-final words, so the technique incorporates a single-word lookahead. We could also repose the task as identification of topic-initial words and avoid the lookahead to have a more on-line process. This is an area for future research.

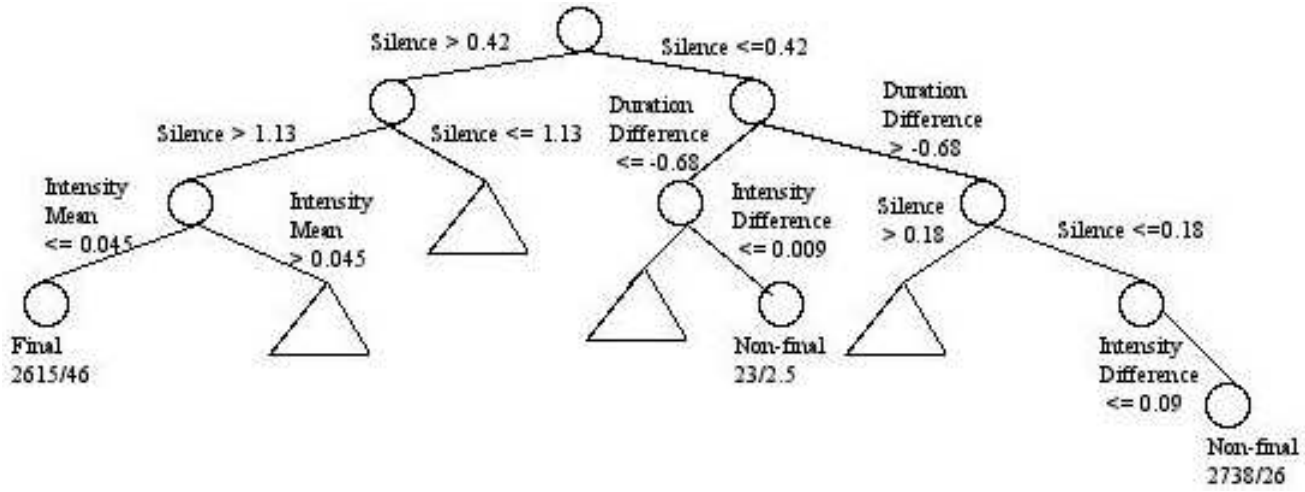

Figure 1: Prosody-based decision tree classifier labeling words as segment-final or non-segment-final

were indicative of such a boundary.[3] These features include the Mandarin Chinese words for "audience", "reporting", and "Voice of America." We used a boolean feature for each such word corresponding to its presence or absence in the current word's environment in the classifier formulation.

7.3 Classifier Training and Testing Configuration

We employed Quinlan's C4.5 (Quinlan, 1992) decision tree classifier to provide a readily interpretable classifier. Now, the vast majority of word positions in our collection are non-segment-final. So, in order to focus training and test on segment boundary identification and to assess the discriminative capability of the classifier, we downsample our corpus to produce a 50/50 split of segment-final and non-final words. We train on 3500 segment-final words[4] and 3500 non-final words, not matched in any way, drawn randomly from the full corpus. We test on a similarly balanced test set of 500 instances.

7.4 Classifier Evaluation

7.4.1 Prosody-only classification

The resulting classifier achieved 95.6% accuracy, with 2% missed boundaries and 7% false alarms. This effectiveness is a substantial improvement over the sample baseline of 50%. A portion of the decision tree is reproduced in Figure 1. Inspection of the tree indicates the key role of silence as well as the use of both contextual and purely local features of pitch, intensity, and duration. The classifier relies on the theoretically and empirically grounded features of pitch, intensity, duration, and silence, where it has been suggested that higher pitch and wider range are associated with topic initiation and lower pitch or narrower range is associated with topic finality.

We performed a set of contrastive experiments to explore the impact of different lexical tones on classification accuracy. We grouped words based on the lexical tone of the initial syllable into high, rising, low, and falling. We found no tone-based differences in classification with all groups achieving 94-96% accuracy. Since the magnitude of the difference in pitch based on discourse position is comparable to that based on lexical tone identity, and the overlap between pitch values in non-final and final positions is relatively small, we obtain consistent results.

7.4.2 Text and Silence-based Classification

In a comparable experiment, we employed only the text similarity, text unigram, and silence duration features to train and test the classifier. These features similarly achieved a 95.6% overall classification accuracy, with 3.6% miss and 5.2% false alarms. Here the best classification accuracy was achieved by the $tf * idf$ weighted 50 word window based text similarity measure. The text unigram features also contributed to overall classifier effectiveness. A portion of the decision tree classifier using text-based features is reproduced in Figure 2.

7.4.3 Combined Prosody and Text Classification

Finally we built a combined classifier integrating all prosodic, textual, and silence features. This classi-

[3] We used the ngram functionality to Boostexter (Schapire and Singer, 2000) to identify these units.

[4] We excluded a small proportion of words for which the pitch tracker returned no results.

fier yielded an accuracy of 96.4%, somewhat better effectiveness, still with more than twice as many false alarms as missed detections. The decision tree utilized all prosodic features. $tf * idf$ weighted cosine similarity alone performed as well as any of the other text similarity or overlap measures. The text unigram features also contributed to overall classifier effectiveness. A portion of the decision tree classifier using prosodic, textual, and silence features is reproduced in Figure 3.

7.5 Feature Comparison

We also performed a set of contrastive experiments with different subsets of available features to assess the dependence on these features.[5] We grouped features into 5 sets: pitch, intensity, duration, silence, and text-similarity. For each of the prosody-only, text-only, and combined prosody and text-based classifiers, we successively removed the feature class at the root of the decision tree and retrained with the remaining features (Table 1).

We observe that although silence duration plays a very significant role in story boundary identification for all feature sets, the richer prosodic and mixed text-prosodic classifiers are much more robust to the absence of silence information. Further we observe that intensity and then pitch play the next most important roles in classification. This behavior can be explained by the observation that, like silence or non-speech regions, pitch and intensity changes provide sharp, local cues to topic finality or initiation. Thus the prosodic features provide some measure of redundancy for the silence feature. In contrast, the text similarity measures apply to relatively wide regions, comparing pairs of 50 word windows.

7.6 Further Feature Integration

Finally we considered effectiveness on a representative test sampling of the full data set, rather than the downsampled, balanced set, adding a proportional number of unseen non-final words to the test set. We observed that although the overall classification accuracy was 95.6% and we were missing only 2% of the story boundaries, we produced a high level of false alarms (4.4%), accounting for most of the observed classification errors. Given the large predominance of non-boundary positions in the real-word distribution, we sought to better understand and reduce this false alarm rate, and hopefully reduce the overall error rates. At the same time, we hoped to avoid a dramatic increase in the miss rate.

[5]For example, VOA Mandarin has been observed stylistically to make idiosyncratically large use of silence at story boundaries. (personal communication, James Allan).

To explore this question, we considered the contribution of each of the three main feature types - prosody, text, and silence - and their combined effects on false alarms. We constructed independent feature-set-specific decision tree classifiers for each of the feature types and compared their independent classifications to those of the integrated classifier. We found that while there was substantial agreement across the different feature-based classifiers in the cases of correct classification, erroneous classifications often occured when the assignment was a minority decision. Specifically, one-third of the false alarms were based on a minority assignemnt, where only the fully integrated classifier deemed the position a boundary or where it agreed with only one of the feature-set-specific classifiers.

Based on these observations, we completed our multi-feature integration by augmenting the decision tree based classification with a voting mechanism. In this configuration, a boundary was only assigned in cases where the integrated classifier agreed with at least two of the feature-set-specific classifiers. This approach reduced the false alarm rate by one-third, to 3.15%, while the miss rate rose only to 2.8%. The overall accuracy on a representative sample distribution reached 96.85%.

8 Conclusion and Future Work

We have demonstrated the utility of prosody-only, text-only, and mixed text-prosody features for automatic topic segmentation of Mandarin Chinese. We have demonstrated the applicability of intonational prosodic features, specifically pitch, intensity, pause and duration, to the identification of topic boundaries in a tone language. We find highly significant decreases in pitch and intensity at topic final positions, and significant increases in word duration. Furthermore, these features in both local form and contextualized form provide the basis for an effective decision tree classifier of boundary positions that does not use term similarity or cue phrase information, but only prosodic features.

We observe similar effectiveness for all feature sets when all features are available, with slightly better classification accuracy for the text and hybrid text-prosody approach. We further observe that the prosody-only and hybrid feature sets are much less sensitive to the absence of individual features, and, in particular, to silence features, as pitch and intensity provide comparable sharp cues to the position of topic boundaries. These findings indicate that prosodic features are robust cues to topic boundaries, both with and without textual cues.

Finally, we demonstrate the joint utility of the dif-

	Prosody-only		Text+Silence		Text+Prosody	
	Accuracy	Pct. Change	Accuracy	Pct. Change	Accuracy	Pct. Change
All	95.6%	0	95.6%	0	96.4%	0
Silence	84.6%	-11.5%	77.4%	-19%	89.6%	-6.9%
Intensity	80.4%	-15.9%			85.4%	-11.4%
Pitch	63.6%	-33.4%			78.6%	-18.5%

Table 1: Reduction in classification accuracy with removal of features. Each row is labeled with the feature that is newly removed from the set of available features.

ferent feature sets - prosodic, textual, and silence. The use of a simple voting mechanism exploits the different contributions of each of the feature-set-specific classifiers in conjunction with the integrated classifier. This final combination allows a substantial reduction of the false alarm rate, reduction in the overall error rate, and only a small increase in the miss rate. Further tuning of relative miss and false alarm rates is certainly possible, but should be tied to a specific task application.

There is still substantial work to be done. We would like to integrate speaker identification for normalization and speaker change detection. We also plan to explore the integration of text and prosodic features for the identification of more fine-grained sub-topic structure, to provide more focused units for information retrieval, summarization, and anaphora resolution. We also plan to explore the interaction of prosodic and textual features with cues from other modalities, such as gaze and gesture, for robust segmentation of varied multi-modal data.

References

D. Beeferman, A. Berger, and J. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34((1-3)):177–210.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

B. Grosz and C. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

M. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.

Julia Hirschberg and Christine Nakatani. 1998. Acoustic indicators of topic segmentation. In *Proceedings on ICSLP-98*.

J. Hirschberg, M. Bacchiani, D. Hindel, P. Isenhour, A. Rosenberg, L. Stark, L. Stead, S. Whittaker, and G. Zamchick. 2001. Scanmail: Browsing and searching speech data by content. In *Proceedings of EUROSPEECH 2001*.

D. Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

C. H. Nakatani, J. Hirschberg, and B. J. Grosz. 1995. Discourse structure in spoken language: Studies on speech corpora. In *Working Notes of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 106–112.

J.R. Quinlan. 1992. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

K. Ross and M. Ostendorf. 1996. Prediction of abstract labels for speech synthesis. *Computer Speech and Language*, 10:155–185.

Gerard Salton. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA.

Robert E. Schapire and Yoram Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39((2/3)):135–168.

X.-N. Shen. 1989. *The Prosody of Mandarin Chinese*, volume 118 of *University of California Publications in Linguistics*. University of California Press.

Marc Swerts. 1997. Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101(1):514–521.

G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57.

C. Wayne. 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Language Resources and Evaluation Conference (LREC) 2000*, pages 1487–1494.

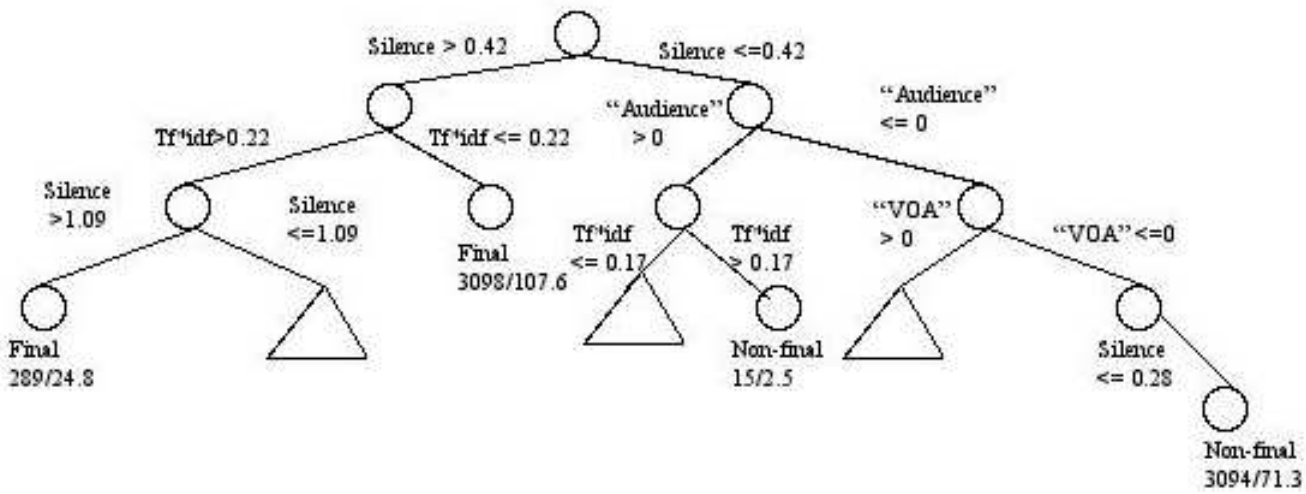

Figure 2: Text-feature-based decision tree classifier labeling words as segment-final or non-segment-final

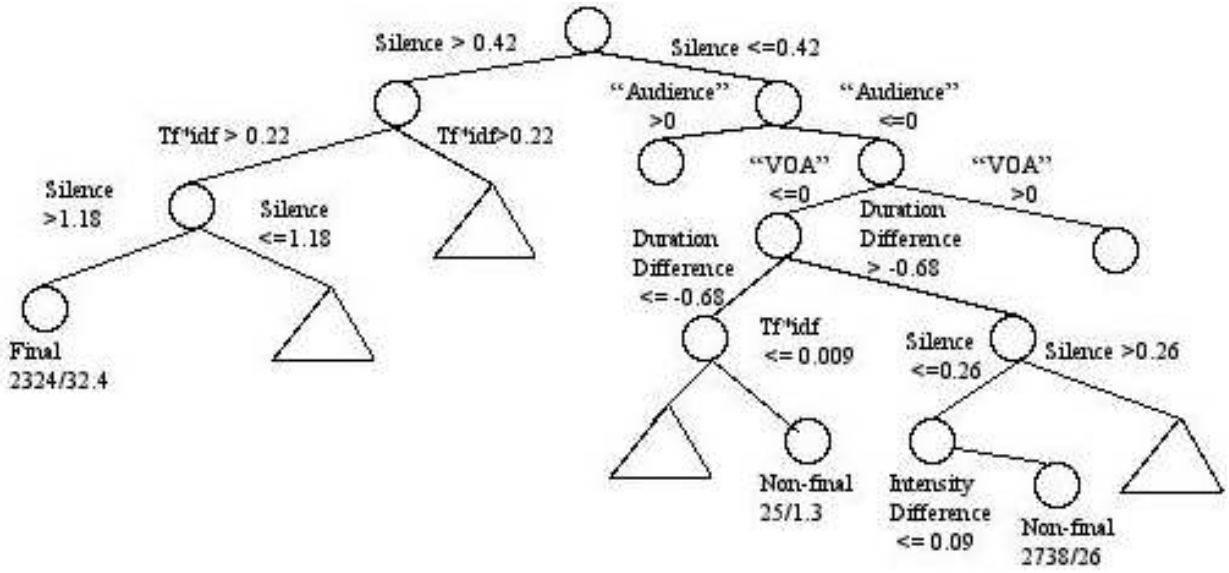

Figure 3: Prosody, text, and silence based decision tree classifier labeling words as segment-final or non-segment-final