Deriving Document Structure from Prosodic Cues

Martin Haase[1], *Werner Kriechbaum*[2], *Gregor Möhler*[1], *Gerhard Stenzel*[2]

[1] Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany
{haasemn,moehler}@ims.uni-stuttgart.de *http://www.ims.uni-stuttgart.de*
[2] IBM Deutschland Entwicklung GmbH, Böblingen, Germany
{kriechba,stenzel}@de.ibm.com *http://www.de.ibm.com/entwicklung/*

Abstract

This study presents an approach for prosody-driven segmentation of speech data. The model is based solely on F_0 contours and RMS envelopes. Phoneme or word information from a speech recognizer is unneccesary. Using data from German broadcast news, we show how this prosodic information can be exploited to retrieve structural information of the spoken text. The suitability of the CART-like algorithm for utterance boundary prediction has been evaluated on 7 five-minutes-news-reports, using 28 reports as training material for the classification tree. Sentence boundaries were predicted with a precision of 93%, at a recall of 88%.

1. Introduction

In this paper we report of an investigation on how structures of written documents and spoken documents correspond to each other. Our goal is to derive the textual structure of a spoken document relying solely on its prosodic parameters. We used German radio news and trained example classification trees based on our data to evaluate how well structural features can be predicted by prosodic features.

There are several means to mark structure in written documents, e.g. using punctuation, paragraphs, headings and different typefaces, to name a few. Spoken documents must reflect this structure in order to provide the same information appropriately to the hearer. In read speech, the main facility to do this is prosodic variation.

Prosody is generally understood to be comprised of several elements, which is variation of pitch and tune, loudness and segmental duration. For an overview, see (Clark & Yallop, 1990). There have been numerous studies relating prosodic structure with document or discourse structure.

Grosz & Hirschberg (1992) investigate how acoustic-prosodic features mark discourse structure in different corpora. More detailed results can also be found in (Hirschberg & Nakatani, 1996), who report on a corpus of direction-giving monologues. They found that phrases in discourse segment-initial places had higher F_0 and RMS values than elsewhere, with shorter subsequent and longer preceding pauses. Segment-final phrases showed the opposite behaviour, combined with a faster speaking rate.

A further substantial part of research has focused mainly on intra-sentential phrasing. Sentence-like utterances typically show a declination in pitch (Clark & Yallop, 1990) (p. 284). F_0 declination can be measured using regression line fits, as reported in (Swerts, Strangert, & Heldner, 1996). Batliner, Weiand, Kießling, & Nöth (1993) compared F_0 offsets in order to classify sentence modality for spontaneous speech.

For modelling of prosodic parameters, see (Schweitzer & Haase, 2000). Their systems predict abstract labels for accents and phrase boundaries based on syntactical structures.

Strom & Widera (1996) examined ‘purely’ prosodic features for prediction of phrase boundaries and accents. They used delexicalized speech and compared the performance of human listeners with an automatic phrase boundary detector. A surprising finding was that the algorithm had only a slightly worse accuracy than human judges.

Shriberg, Stolcke, Hakkani-Tür, & Tür (2000) describe a work similar to ours. Besides ‘purely’ prosodic features, they include also speech recognizer transcripts to build a language model and combine it with the prosodic model. The trained classi-

fication tree can automatically distinguish topic and sentence boundaries.

2. Method

2.1. Data

We used a new corpus of broadcast news from the German radio station 'Südwestdeutscher Rundfunk' (SWR). It consists of 166 minutes of recorded speech, which divides into 35 news reports, approx. 5 minutes each. There are 7 different male speakers. Every report is read by one speaker.

A report consists of several topic blocks. Each topic has a heading - a city name -, followed by one or more sentences forming its content. Pitch and intensity were extracted using the program `Praat`[1]. Text alignment of the training data sets was done using the word-level `TALC`™ Aligner[2].

2.2. Models for Prosodic Features

Two restrictions apply to the present segmentation approach. (1) It must be 'purely' prosodic, as we do not want to employ a speech recognizer. Consequently, there is no word, syllable, or phoneme information available. (2) It should work 'online', scanning F_0 and RMS curves of the speech signal with a window of predefined length.

For the determination of the window, we followed a somewhat unorthodox approach: Features are calculated for all *voiceless regions* in the signal. A minimum threshold for the length of such regions (100 ms) allows for an initial distinction between potential pauses and unvoiced consonants. In figure 1, some of the extracted features are drawn.

A first set of features can be computed directly from the F_0 and RMS curves. It consists of the length `w` of voiceless regions `X` (corr. to pause length), F_0 Off- and Onset `f1` and `f2`, and their difference, indicating a relative change in F_0. The feature `mi` is the mean over intensity values lying within `X`.

Furthermore, regression lines have been calculated over the F_0 values preceding and following any voiceless region `X`. They are intended to reflect the phenomenon of F_0 declination, cf. (Swerts et al., 1996). We chose fixed length windows of 500, 1000, 2000 and 4000 ms, starting from the margins of `X`. The according slopes are reflected by the features `s1` and `s2` (in Hz/s). We also included estimated values for F_0 Off- and Onset `v1` and `v2`, and their difference.

Other features comprise mean and median values over adjacent F_0 values surrounding `X` to reflect the pitch register, and the respective standard deviation and interquartile distance to model pitch range.

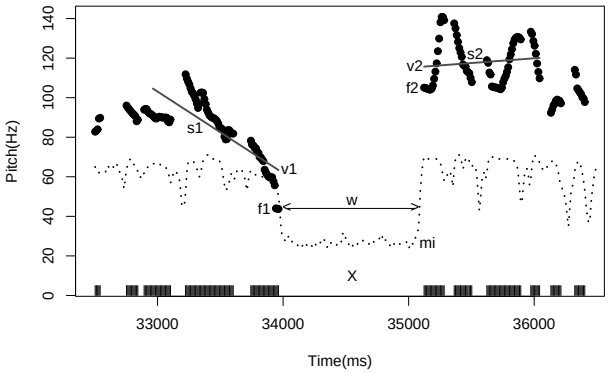

Figure 1: Sample speech F_0 data illustrating some of the extracted features. Also shown: RMS values (in dB, dotted line) and voiced information (bottom). Features described belong to the voiceless region `X`.

3. Data Exploration

The feature space spanned by the described features is rather complex. As a first glance, we show our data with only two important features (`w` and `s1`) in figure 2. Each datapoint corresponds to a voiceless region ($>$ 100 ms). Only the datapoints with a non-ambiguous word alignment are shown.

Some interesting findings are revealed by visual inspection:

- Topic block headings (city names) show a very steep declination line.[3]
- Sentence and paragraph ends do show a slightly steeper falling slope than other elements, but the prime distinction is based on pause length.
- Pause length after paragraph breaks is longer than after sentences.

[1] Many thanks to Paul Boersma for a copy of `Praat`.

[2] The IBM Deutschland Entwicklung GmbH supported this project with these and further resources.

[3] This is caused by the pause preceding a topic heading. The regression line is calculated for only one word, thus not measuring global declination but the local final fall.

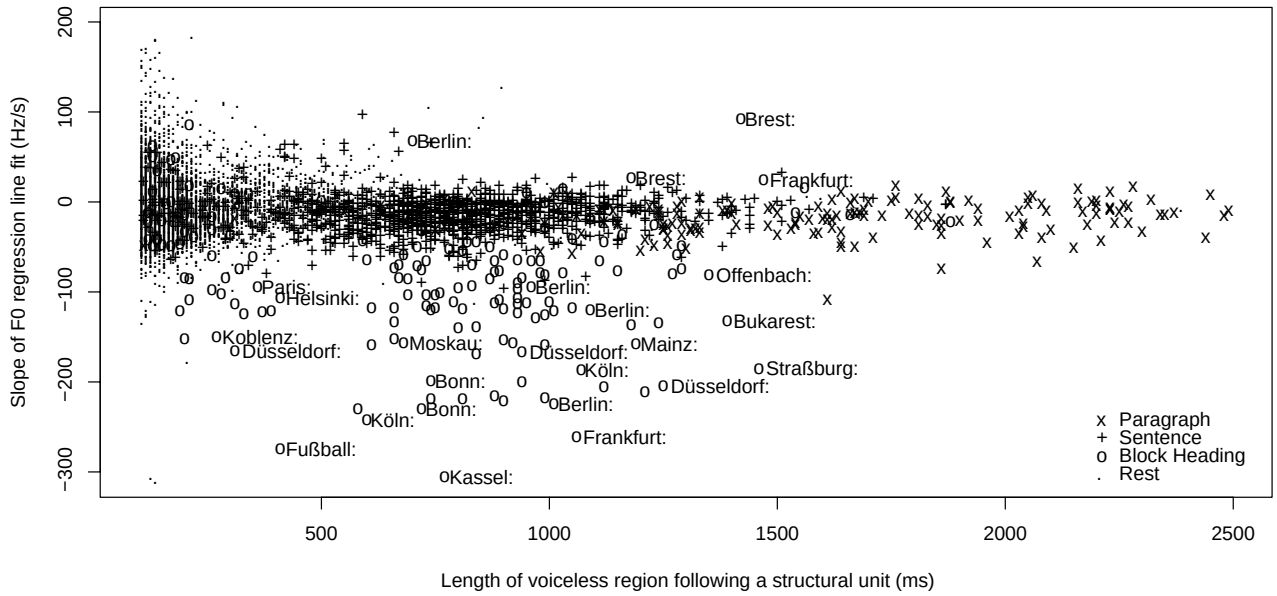

Figure 2: SWR data: Length of voiceless regions X and slopes of regression lines over F_0-values in a window of 1000 ms preceding X. The shape of data points reflects their structural function. Some points have the corresponding word annotated.

The experimental results described in the next section will confirm these assumptions.

4. Prediction of Document Structure

For the prediction task, we use classification trees as desribed in (Ripley, 1996). F_0 curves in each report were normalized with a z-score transformation.[4]

For evaluation, we randomly selected one fifth of our data (7 reports) as a test corpus. The remaining 28 reports were used as training data. In a first training frame we grew classification trees for the distinction between utterance boundaries and non-boundaries.

Pause length turns out to be the most important feature for tree construction. The simplest 'tree' with two leaves chooses a value of w=475 ms for distinguishing utterance boundaries from non-boundaries. This confirms the expectations from figure 2. The error rate[5] of this binary classifier is 0.048, (precision/recall[6] 0.85/0.84). The tree with the best results has 15 leaves. Its error rate decreases to 0.032 (precision/recall 0.91/0.88).

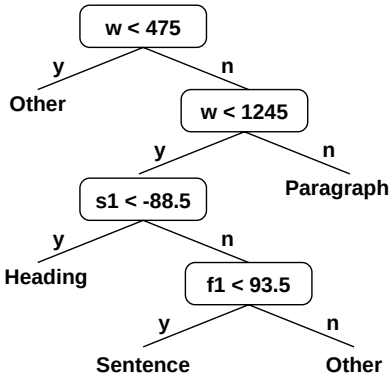

Figure 3: Simple tree for structure prediction

In a second training frame the distinction between the different kinds of boundaries is made, i.e. sentence, paragraph, heading and other. A simple tree from this training frame is depicted in figure 3. As expected, slope information is used to distinguish block headings from other types. F_0 offset turns out to be another important criterion besides pause

[4] Using mean and standard deviation computed over a whole report. For a realtime version, estimated values could be computed from the first 20 seconds of a report.

[5] We use the metric from (Fiscus, Doddington, Garofolo, & Martin, 1999). Chance probability in our corpus differs from the TDT2 numbers: only 15% of all voiceless regions X with a non-ambiguous word alignment are utterance boundaries. Thus weights for misses and false alarms are 0.15 and 0.85, resp.

[6] See (Manning & Schütze, 1999) for details.

length. The confusion matrix of this tree is given in table 1. The numbers correspond to an error rate of 0.038 (precision/recall 0.93/0.81).[7]

	Para	Sent	Head	Other
Paragraph	27	15	3	2
Sentence	8	185	7	19
Heading	0	6	12	0
Other	2	47	12	1726

Table 1: Evaluation of the classification tree from figure 3 on the test corpus. Columns are reference classes, rows predicted.

When the growth of the tree is stopped later, classification results improve further. The optimal tree for our data has 25 leaves. Its confusion matrix is given in table 2. The results show that the different utterance types can be predicted quite reliably. The error rate decreases still slightly to 0.030, with precision/recall at 0.93/0.88, respectively.

	Para	Sent	Head	Other
Paragraph	28	3	1	1
Sentence	7	212	7	19
Heading	0	4	22	3
Other	2	34	4	1724

Table 2: Evaluation of the optimal classification tree.

5. Discussion

Several extensions to this approach are conceivable. The prosodic feature set could be enlarged by segmental durations. Another promising idea is the combination with a language model as reported in (Shriberg et al., 2000).

The results gained in this study are quite encouraging towards other applications in the domain of audio segmentation and structure prediction. The German broadcast news corpus used in this experiment constitues a somewhat 'ideal' material for this task. It has a very formal style with speakers realizing a controlled intonation. Major emphasis lies on a clear communication of structural information. Other speech data may or may not be equally suited for this approach.

[7] For the evaluation metrics, the 3 utterance types (Paragraph/Sentence/Heading) were merged into one event.

6. References

Batliner, A., Weiand, C., Kießling, A., & Nöth, E. (1993). Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody. In *Proc. ESCA Workshop on prosody, Lund*.

Clark, J. & Yallop, C. (1990). *An Introduction to Phonetics and Phonology*. Basil Blackwell, Oxford/Cambridge.

Fiscus, J. G., Doddington, G., Garofolo, J. S., & Martin, A. (1999). NIST's 1998 Topic Detection and Tracking Evaluation (TDT2). In *Proceedings of the DARPA Broadcast News Workshop.* Also see http://www.nist.gov/speech/tests/tdt/tdt98/index.htm.

Grosz, B. & Hirschberg, J. (1992). Intonational Characteristics of Discourse Structure. In Ohala (Ed.), *Proc. 2nd International Conference on Spoken Language Processing, Banff*, Vol. 1, pp. 429–432.

Hirschberg, J. & Nakatani, C. H. (1996). A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues. In *ACL '96 proceedings*.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Schweitzer, A. & Haase, M. (2000). Zwei Ansätze zur syntaxgesteuerten Prosodiegenerierung. In *Tagungsband der KONVENS 2000 - Sprachkommunikation*. VDE-Verlag, Berlin.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., & Tür, G. (2000). Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. *Speech Communication*, *32*(1-2).

Strom, V. & Widera, C. (1996). What's in the 'Pure' Prosody?. In *Proc. ICSLP 96, Philadelphia*. Also available as Verbmobil report 168.

Swerts, M., Strangert, E., & Heldner, M. (1996). F_0 Declination in Read-aloud and Spontaneous Speech. In *Proceedings 4th Int. Conf. on Spoken Language Processing, Philadelphia*.