Automatic Sentence Boundary Detection for German Broadcast News

Georgi Dzhambazov, Rolf Bardeli

Fraunhofer Institute for Intelligent Analysis and Information Systems,
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
Email: {georgi.dzhambazov,rolf.bardeli}@iais.fraunhofer.de
Web: www.iais.fraunhofer.de

Abstract

In this work we aim at enriching the transcript of an automatic speech recognition system with punctuation by automatically detecting sentence ends. We make use of a simple word-based language model and combine it with a decision tree for the acoustic features of speech. The focus lies on selecting robust acoustic features that reflect the prosodic characteristics of the German language in a most optimal way. We arrive at a Sentence Unit Error Rate of 54 compared to the state-of-the art rate for English of 61, by applying a comparable detection system. This is a sound indication that prosody has a stronger cue on perception of sentence boundaries for German than for English. Our work is, to our knowledge, the first system developed for sentence boundary detection for the broadcast news domain for German language. Our results can therefore serve as a baseline for further studies in this scenario.

1 Introduction

Unlike written text, in which punctuation is present in most application scenarios, automatically-generated transcripts of speech lack any indications about structuring into phrases and sentences. Automatic segmentation of spoken language into sentence units has numerous applications within the domain of natural language processing. Having a sentence-wise segmented transcript is for example a crucial prerequisite for tasks such as speech understanding or machine translation.

In this work we aim at enriching the transcript of an automatic speech recognition (ASR) system with punctuation for sentence ends. We address the task of automatic sentence boundary detection (ASBD) for broadcast news speech in German. Speech presents to the listener two informational sources: a lexical stream consisting of a sequence of words and an acoustic stream with prosodic phenomena. Prosody is an umbrella term that covers the acoustic perceptual cues of speech like pauses, duration, intonation, stress.

Although the detection scheme we use has been first proposed and tested against an English corpus [4], this is, to our knowledge, the first work that applies it for German.

2 Related Work

In recent years approaches combining textual and prosodic information sources have proved to produce the best accuracy [4]. In that work an extensive set of prosodic features has been proposed and their performance evaluated on an English corpus from the broadcast news domain. The modeling of text is handled by a word-based 4-gram language model, which is additionally aware of sentence boundaries. A more recent system which builds upon that by integrating a more sophisticated language model is presented in [1]. The performance of this system will be taken as a reference in this work.

There has been very little amount of work on ASBD for German. A set of prosodic features comparable to ours has been applied by [3] for detecting prosodically marked boundaries in spontaneous speech. A system that recognizes sentence modality is presented in [7]. It is based on training different HMMs for different sentence modalities. The model captures a template for the prosodic feature contours peculiar for a given sentence type. This work was tested on a corpus of read German prose. No comparison with it is however possible since no results on the ASBD part of the task were reported for German.

3 Approach

3.1 General Setting

We will use the following notation:

w_t the word at position t in the text transcript — be it manually-generated or hypothesized by the ASR.

f_t Prosody feature vector extracted from a time window around the beginning of the tth word.

e_t the type of sentence boundary preceding the tth word

We model sentence boundaries as hidden events preceding each word in the text transcript. In other words, we perceive the word stream as a sequence of pairs each consisting of one word following a hidden event, e.g. the pair at time t is $\langle e_t, w_t \rangle$. For each event at given time t there exist two possibilities: $e_t \in \{< s >, < n >\}$. These correspond to sentence boundary $< s >$ and non-sentence boundary $< n >$.

On training, the speech utterances are aligned to the reference text transcript, which provides a set of starting and ending timestamps on phoneme and word level. This temporal information enables the extraction of the prosodic features from the speech signal. On detection the timestamps are available from the ASR module.

Our goal is to determine the optimal event sequence $\hat{E}$, given the observed word sequence W and the acoustic feature vector sequence F:

$$\begin{aligned}\hat{E} &= \arg\max_E P(E|W,F) \\ &\approx \arg\max_E P(W,E)P(F|W,E)\end{aligned}$$

Here the term $P(W,E)$ can be modeled by a well-suited language model, and $P(F|W,E)$ corresponds to a prosodic model. This equation is adopted from [1].

3.2 Classification Approach

To integrate the two models we adopt the approach of [1] to use a first order hidden Markov model (HMM) that has two states: $< s >$ and $< n >$ and is ergodic in its topology.

Further, each state emits two tokens — the current word identity w_t and a feature vector f_t. These account for the two parallel observed information streams W and F.

3.2.1 Language Model

To model the joint probability of words and events $P(W,E)$ we apply the hidden event language model (HELM) as introduced by [5]. It is an extension of a traditional statistical n-gram model, enriched with information about tokens for sentence boundaries.

On training, the events are explicitly modeled. More precisely, the event preceding the current word and the event preceding the previous word are incorporated in the model. These correspond to the current state and the previous state, when decoding using the HMM. This allows to apply the posterior probabilities from the HELM as transition probabilities in the HMM. The events are called *hidden* since on detection they are not present in the word sequence and have to be recognized.

Since language modeling was not a focus of this work, we opted for a simple word-based trigram HELM interpolated from three different training corpora from printed news.

3.2.2 Prosodic Model

The prosodic model has to model the probability of observing an emitted prosodic feature vector being in a given state $P(f_t|e_t,W)$. To approximate this posterior probability a classification decision tree is trained. Feature vectors f_t are extracted from the training corpus, in which the corresponding events e_t are annotated. As this is an observation probability, it can therefore serve as the emission term in the HMM.

This probability is weakly conditioned on the word-sequence in the vicinity of the prosodic feature at time t. This is due to the fact that some of the features depend on the word-sequence. This dependency however can be eliminated by normalizing the corresponding dependent feature. For example the Last Syllable Duration (3.3.2) is normalized with respect to the number of phonemes in the syllable making it thus word-independent. Now applying Bayes Rule this becomes:

$$P(f_t|e_t) = \frac{P(e_t|f_t)P(f_t)}{P(e_t)}$$

Since the probability $P(f_t)$ is fixed for the two cases $e_t =< s >$ and $e_t =< n >$, it can be ignored on comparison of the two classes.

One idea for $P(e_t)$ can be to approximate it offline based on the average length of a sentence. Another idea, that has been proposed by [4], is to downsample the training set to make $P(e_t =< s >) = P(e_t =< n >) = 1$. We have tested both of the approaches choosing a non-downsampled version in the end. On recognition, the posteriors $P(e_t|f_t)$ at each tth word boundary can be read off the leaves of the decision tree.

A scheme of the integration of the two models into the HMM is presented in Figure 1.

3.3 Prosodic Features

German and English are in the same linguistic family and are thus similar in the way prosody is expressed. In this respect we adopt some of the prosodic descriptors which proved to be efficient for English. We have picked three of the four feature groups that showed the highest discriminatory power for English according to the evaluation with the *feature usage* metric [4]. To have an intuition for the applicability of the selected features to German, we analyzed the statistical distribution for each feature. All prosodic features are extracted from a window around the word boundary preceding each word and are fed into a classification decision tree.

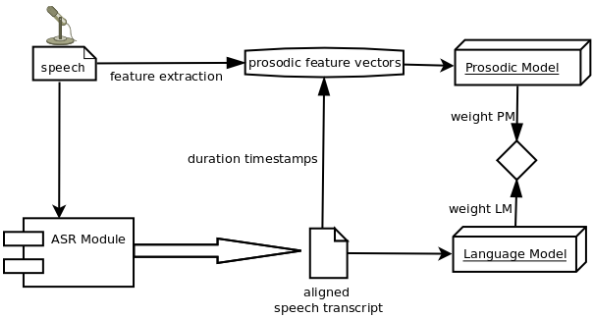

Figure 1: System architecture. The prosodic model and the language model are integrated using a HMM by corresponding model weights. The prosodic model is built on features that are extracted from the speech signal and have timestamps of the corresponding words from the aligned speech transcript, which is output from the ASR module. The language model is built directly from the speech transcript.

3.3.1 Pause Duration

This feature represents the duration of silent pauses between words. Unlike in conversational unprepared speech, speakers of news journals make emphasized use of pauses between words to demarcate semantic phrases. This fact has been proved by the recall of 56.9 of the prosodic classifier trained only with pause duration compared to the recall of 62.3 of all prosodic features together.

3.3.2 Last Syllable Duration

The last syllable duration accounts for the so called pre-boundary lengthening phenomenon — lengthening of the last word syllable preceding a sentence boundary. We take the average phone duration to normalize with respect to the number of phonemes in the syllable. It is defined like this:

$$avrgPhonemeDur_t = \frac{\sum_p NormalizedPhonemeDuration_p}{numberPhonemesInRhyme}$$

where p is an index for the given phoneme. Additionally, the duration for each phoneme is normalized to compensate for the different speaking rate of speakers. Our analysis proved a pronounced tendency of news readers to lengthen the last word syllable preceding a sentence boundary. This is confirmed by the shape of the histogram for the two classes in the training set as can be seen in Figure 2.

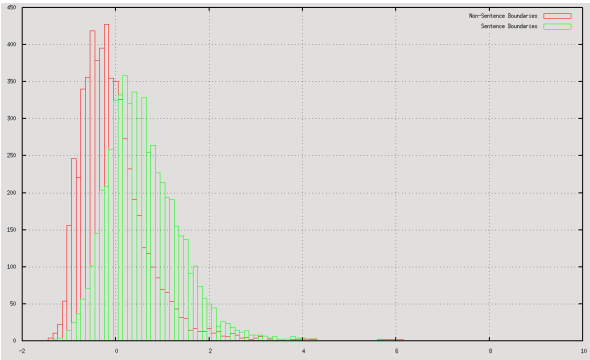

Figure 2: Histogram on the feature *last syllable duration* on training set. Green represents the class of sentence boundaries and red non-sentence boundaries

3.3.3 Fundamental Frequency Reset

Intonation, among its other roles, serves as well the purpose to outline the transition to new sentences by resetting the pitch at the beginning of a new sentence. This creates the feeling for intonational break since intonation is otherwise relatively continuous within an intonational phrase. Raw fundamental frequency (f0) tracks were extracted from speech using the pitch extraction algorithm of Talkin [6]. Since f0 does not make sense for non-voiced intervals of words, frames belonging to such intervals were simply disregarded. Further smoothing with median filtering of size 7 frames proved to improve the feature.

In order to model the reset of f0 preceding word w_t one f0 value is needed, which is representative for the word preceding the word boundary $f0Repr_{t-1}$, and one for the word following it $f0Repr_t$. As a representative value $f0Repr_{t-1}$ the median of a fixed time window of voiced values is taken. Similarly, for $f0Repr_t$ the median of a window of same size at the beginning of the word is taken. This way any outlier values can be avoided, which occur at word beginnings due to voice onsets and at word endings due to non-modal voicing. We have explored several different f0-based features, proposed by [4] which model this f0 reset. The one selected as most discriminating is a variation of the *reset feature* and can be expressed this way:

$$f0Diff = \frac{|\log(f0Repr_t) - \log(f0Repr_{t-1}) + offsetConst|}{\log(f0Repr_{t-1})}$$

Our modification is adding the absolute value, which is motivated by the fact that the histogram of the feature is relatively symmetric and folding it in an appropriate way would increase the discriminatory power of the feature. To account for the optimal axis of folding we have introduced a constant offset term $offsetConst$.

Notably, the inclusion of this feature in the decision tree improves recall from 57.5 to 62.3.

The final decision tree structure can be seen in Figure 3. Note that the pause-based feature is at the highest level, followed by the f0-based feature. The phoneme-based feature has the smallest contribution being lowest in the tree hierarchy. This precedence resembles the one of the tree for English presented in [4], which confirms the similarity between German and English with respect to the role prosodic features play for demarcating sentence breaks.

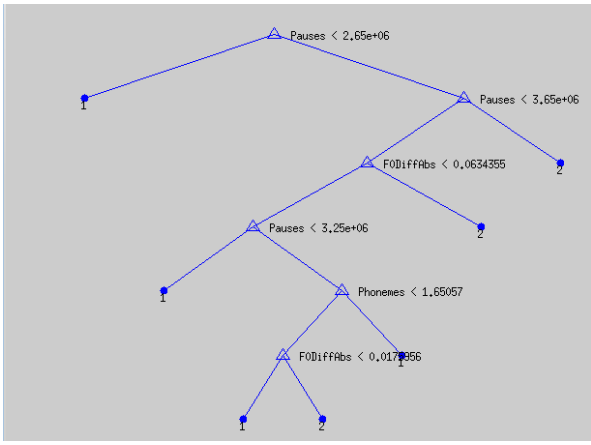

Figure 3: A binary classification decision tree trained on the three prosodical features. Nodes present the decision rules (questions). The left branch corresponds to "yes". On the leaves the label 1 stands for non-sentence boundary and 2 for sentence boundary. This is the tree used in the final system and has SuER = 58.0; precision = 77.8; recall = 58.7.

3.4 Experimental Setup

The decision tree is trained on prosodic feature vectors extracted around word boundaries of a training corpus of 100K words. Testing is run on a test corpus of 15K words, both corpora based on three German TV news journals. Since the complete structure of our classification approach has been adopted from [1] we chose to use this system as a reference. The differences in our work lie in the non-downsampled training prosodic feature set and the modification of the Fundamental Frequency Reset feature.

Since the word error rate of an ASR varies from one system to another and from one language to another, we conduct experiments with manually-generated reference word transcripts. This allows us to have an unbiased basis for comparison with other segmentation systems and to reflect the performance of the prosodic module beyond perfect word information.

As a decoding algorithm we have chosen the Forward-Backward algorithm based on the HMM (3.2) which combines the language model and the prosodic model. Forward-Backward is used as well in [1].

For a test speech excerpt, the detection of the boundary events is done by running the decoding algorithm for each event e_t in turn, hypothesizing a $< s >$ or a $< n >$. The detection process involves applying the algorithm sequentially for each next word boundary by shifting the context window one word to the right, selecting the event having higher posterior probability.

4 Results and Discussion

We measure accuracy by means of the metric Sentence Unit Error Rate (SuER) imposed by the National Institute of Standards and Technology (NIST) as part of their rich transcription evaluation task [2].

Models	SuER	P	R	F
LM only	91.1	66.4	18.0	28.3
PM only	58.0	77.8	58.7	66.9
PM & LM	54.3	85.4	55.2	68.0

Table 1: Evaluation results on reference transcripts
PM = Prosodic Model; LM = Language Model; P = precision; R = recall; F = F-measure

SuER is defined as

$$SuER = \frac{M+S}{C+M}$$

where C, M, S denote respectively the count of correct, missing and spurious identifications of sentence units.

To be able to compare with other systems, we use as well the F-measure combining precision and recall

$$F = \frac{2PR}{P+R}$$

in which complying with the above-mentioned notation precision is defined as $P = \frac{C}{C+S}$ and recall as $R = \frac{C}{C+M}$. Note that for compatibility with precision and recall, we use in this work SuER predominantly in percent scale.

Table 1 presents the SuER, precision recall and the respective F-measure of the final system evaluated on a reference transcript. The performance of the HMM with only one of the models is measured separately, whereas the last row stands for the combination of both models. The model weights for the combination, which result in minimal error metric score, were found empirically to be 0.7 for the language model and 0.3 for the prosodic model.

It can be inferred from the table that the good precision of the LM and the relatively good recall of the PM complement each other to arrive at a reasonable final SuER.

The Forward-Backward decoding allows us to empirically adjust the sizes of the forward window $\ell = 6$ and backward window $m = 3$. These numbers are in accordance with the intuitive expectation that the history word sequence has more pronounced influence on the sentence end since it accounts for its meaning. A length of six corresponds roughly to the average length of a simple clause. On the other hand, some specific word or phrases at the beginning of the word sequence, following the current hypothesized event, can signal the commencement to a new thought. Usually such phrases are not longer than three words.

To compare our accuracy with [1] we denote this system consisting of only a 4-gram LM for broadcast domain by *Liu: B word-LM*. The same system combined with a prosodic decision tree for the broadcast domain and for the conversational domain are named respectively *Liu: B word-LM DT* and *Liu: C word-LM DT* . Table 2 systematizes the the performance of these systems. We arrive at a 7 percent absolute improvement of the SuER compared to *Liu: B word-LM DT*, despite the fact that our LM model alone scores significantly worse than theirs.

A very similar set of prosodic features but with a multilayer perceptron classifier has been applied for conversational speech in German in [3]. We denote this approach as *Noeth: C MLP*. They have additionally integrated a stochastic language model comparable to ours, which however has varying n-gram length and divides words into categorical classes. This approach is denoted by *Noeth: C LM MLP*. It can be seen that *Noeth: C MLP* achieves accuracy almost twice worse than our PM-only system, which is supposedly due to the inferiority of MLPs to decision trees for this task. Most importantly, we achieve similar result to the system *Noeth: C LM MLP*, although ASBD for the case of conversational speech is expected to score significantly better than for broadcast news. This is due to the fact that sentences are mainly simple clauses, unlike in broadcast news. This statement is supported for English by the twice better accuracy of *Liu: word-LM DT* compared to *Liu: C word-LM DT*.

System	SuER
Liu: B word-LM	73.7
Liu: B word-LM DT	61.4
Liu: C word-LM DT	33.2
Noeth: C MLP	102.4
Noeth: C LM MLP	52.7

Table 2: SuER scores on reference transcripts for other systems

5 Conclusion

In this work we apply one of the best-performing ASBD approaches to German broadcast news. It yields significantly better detection rate than for English, even though we make use of a very simple LM. Results of the comparison with hitherto systems let us conclude that one reason for that might be that the discriminating power of prosody for German broadcast news is stronger than for English.

Furthermore, being the only ASBD system for broadcast news in German, we hope it can serve as a baseline work in further applications based on prosodic features.

References

[1] Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary P. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1526–1540, 2006.

[2] NIST. Fall 2004 rich transcription (rt-04f) evaluation plan, October 2004.

[3] Elmar Nöth, Anton Batliner, Andreas Kießling, Ralf Kompe, and Heinrich Niemann. Verbmobil: The use of prosody in the linguistic components of a speech understanding system. *Speech and Audio Processing, IEEE Transactions on*, 8(5):519–532, 2000.

[4] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, 32(1):127–154, 2000.

[5] Andreas Stolcke and Elizabeth Shriberg. Automatic linguistic segmentation of conversational speech. In *Proc. ICSLP*, pages 1005–1008, 1996.

[6] D. Talkin. A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, 495:518, 1995.

[7] K. Vicsi and G. Szaszák. Using prosody to improve automatic speech recognition. *Speech Communication*, 52(5):413–426, 2010.