Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Statistical methods for the automatic labelling of German prosody

Michael Lehning

TU Braunschweig

Vm

Report 159
August 13, 1996

August 13, 1996

Michael Lehning

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstr. 22
38092 Braunschweig

Tel.: (0531) 391 - 2476
Fax: (0531) 391 - 8218
e-mail: lehning@ifn.ing.tu-bs.de

Gehört zum Antragsabschnitt: 14.3 Werkzeuge zur prosodischen Etikettierung

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 N 0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Contents

1 Introduction

It is well established that the use of prosodic knowledge can improve speech recognition and understanding. Nevertheless the prosodic cues (fundamental frequency, energy) are frequently ignored in the implementation of such systems. Recent advances in speech technology, requiring the collection of large speech corpora for analysis and training, have placed an increasing emphasis on the annotation of speech. In order to provide useful databases for speech analysis, large volumes of natural speech must be both prosodically labelled and annotated.

In the German compound project VERBMOBIL our institute's task is to provide prosodically labelled data for the German speech community.

In the last two years we have prosodically labelled hundreds of utterances (one hour of spontaneous speech) [8].

To support the human transcriber with this task we have developped a semi-automatic method.

The basic idea of our approach is to use statistical methods for the generation of a prototypical prosodic description. The prosodic description includes the word boundaries in the speech signal, the accents, and the phrase boundaries. This prototypical information about the prosody can be integrated into the human labelling strategy. It is considered to be easier to *verify* the predicted description than to *label* this data without this information.

2 Outline of the system

We have developed a modular system. Interfaces have been defined for the interaction between all modules. So we are able to substitute each module in order to try out other algorithms and knowledge sources.

For generating a prototypical description we fall back upon the sampled signal and the text of the utterance and various knowledge sources:

Knowledge source	Usage
pronunciation lexicon	Conversion into the phonetic description
HMM phonetic units	Alignment of speech signal and phonetic description
categorical lexicon	Conversion into grammatical units
categorical n-gram	Prediction of accents and phrase boundaries

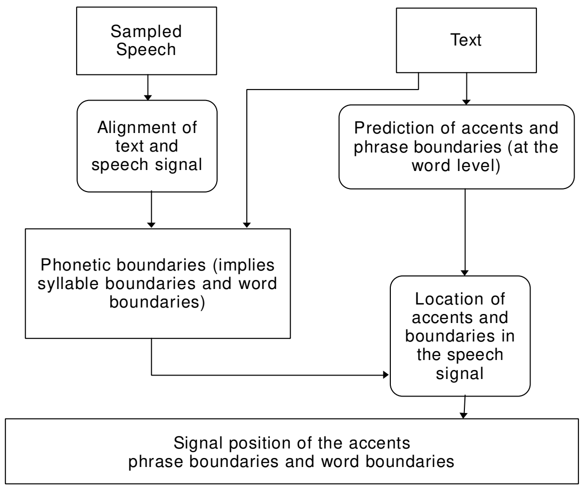


Fig. 1: Outline of the system

The entire procedure comprises three steps:

1. automatic Segmentation of the speech signal into phonetic units (this implies word boundaries),
2. automatic Prediction of phrase boundaries and accents from the text of the utterance,
3. location of boundaries and accents within the speech signal by combining the results from step 1 and step 2.

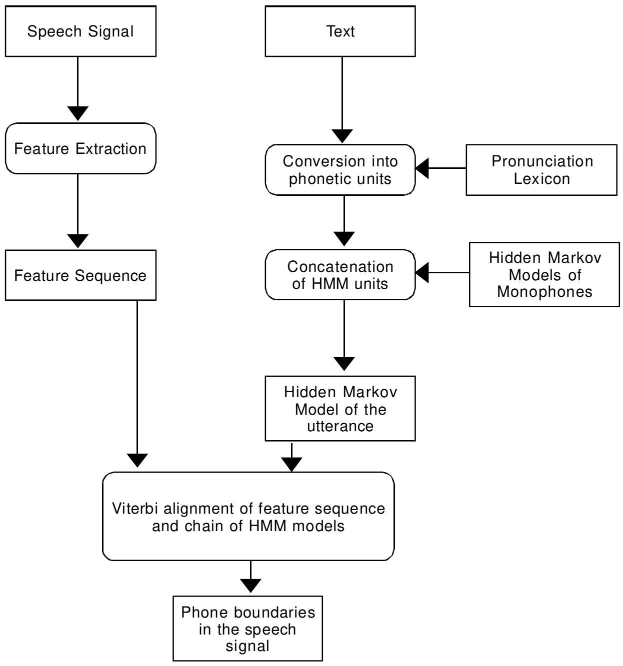


Fig. 2: Detection of phonetic boundaries

3 Alignment of text and speech signal

We have developed an automatic procedure for the segmentation of speech based on an HMM recognition system:

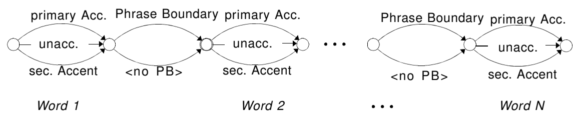
From the text the phonetic string of the utterance is generated (pronunciation lexicon). This is then translated into a Hidden Markov Model of the utterance by the concatenation of the HMMs for each phonetic unit. The Viterbi algorithm is then used to align the short time frequency representation of the speech signal to the HMM of the utterance. This yields the segmentation of the speech signal into phonetic units [3],[5]. In the pronunciation lexicon each phonetic representation of a word comes with its syllable boundaries. So we can extract the syllable boundaries from the phonetic boundaries.

We use a semicontinuous Hidden Markov Model with 47 phonetic units without context information.

4 Prediction of phrase boundaries and accents

For the prediction of phrase boundaries and accents we convert the text of the utterance into a string of grammatical units (with the use of the grammar lexicon) and start a suboptimal graph search procedure (beam search) to find the optimal path through a graph as shown below.

The score for each possible path is calculated by a categorical trigram model [2],[6].



Prediction of phrase accents
and boundaries by a graph search

This optimal path yields accents and phrase boundaries within the utterance. The trigram model is based on grammatical categories (e.g. noun, verb) augmented by the accent of the corresponding word: noun (non accentuated noun), noun_PA (noun carrying primary accent), noun_SA (noun carrying secondary accent), and so on [2].

This categorical trigram model was trained on a a set of prosodically labelled data. Each word of text was converted into its grammatical category (e.g. noun, verb) via the grammar lexicon, and the accent was taken from the labelled data.

Phrase boundaries are handled as a special word group (|) (see example below) Example of the prediction of accents and phrase boundaries:

```
Transliteration:
  dann w"urde ich sagen da"s wir
  noch ein' Termin
  ausmachen  <breathing>
```

```
Conversion into grammatical units:
  adv verb pro verb konj pro
  adv det noun
  verb <breathing>

Results of the prediction:
  adv verb pro verb_PA |  konj pro
  adv det  noun_SA
  verb | <breathing>

Conversion to the text level:
  dann w"urde ich sagen_<PA> |  da"s wir
  noch ein' Termin_<PA>
  ausmachen | <breathing>
```

4.1 Implementation of the beam search algorithm

For the prediction of phrase boundaries and accents we calculate the score fore each possible path in the following way (see flow chart):

Before starting the beam search we convert the text (N words) into the sequence of N symbols (grammatical units) $S_1 \dots S_N$ (see example above). The probabilities are calculated by the categorical trigram.

We have chosen a categorical language model for two reasons:

- small amount of training data (600 utterancens) but a lexicon size of about 2000 words. So we have seen a lot of words only for a few times in the training material;

- there is a good correspondence of prosodic and syntactic boundaries [1]. So we have decided to train the language model on the syntactic word categories.

- Initialize history buffer $HB(1)$
- FOR $1 \leq n < N$
  - Initialize $P(H_{best})$
  - FOR each history $h = h_1 \ldots h_K$ in the history buffer $HB(n)$
    - expand history h into 6 new histories
    - Case 1: S_{n+1} carries no accent
    - $H^1 = H + S_{n+1}$ and
    - $H^2 = H + P + S_{n+1}$
    - Case 2: S_{n+1} carries secondary accent
    - $H^3 = H + S_{n+1}$ and
    - $H^4 = H + P + S_{n+1}$
    - Case 3: S_{n+1} carries primay accent
    - $H^5 = H + S_{n+1}$ and
    - $H^6 = H + P + S_{n+1}$
    - FOR $i = 1, 3, 5$
      - Calculate $P(H^i) = P(H) + p(S_{n+1}|h_{K-1}h_K)$
    - FOR $i = 2, 4, 6$
      - Calculate $P(H^i) = P(H) + p(S_{n+1}|h_K P) + p(P|h_{K-1}h_K)$
    - FOR $i = 1, \ldots 6$
      - IF $P(H^i) > P(H_{best})$
      - THEN $P(H_{best}) = P(H^i)$
      - IF $P(H^i) > P(H_{best}) * \Theta$
      - THEN
        - Store H^i in the history buffer HB(n+1)
- extract boundaries and accents from the best history

5 Location of accents and boundaries in the speech signal

The signal position of the accent is found by a two step algorithm:

1. For all accentuated words retrieve the syllable carrying the lexical stress from the phonetic lexicon.

2. Locate the center of this syllable in the speech signal from the phonetic time alignment.

The predicted phrase boundaries are located at the beginning of the first word and the end of the last word in the phrase.

This automatic preclassification will then be manually corrected.

The prediction of phrase boundaries and accents from the text can also be used for the controlling of prosodic parameters for speech synthesis systems.

6 Results

The prediction of accents and phrase boundaries was compared with the manual labelling for 397 words (21 utterances). The comparison yields the following confusion matrices:

Accents		
class	classified as	
	unaccentuated	accentuated
unaccentuated	277	46
accentuated	18	56

Correct classification: 83.88 %

Phrase boundaries		
class	classified as	
	no boundary	boundary
no boundary	297	18
boundary	20	62

Correct classification: 90.43 %

The percentages in these tables show a satisfying correspondence for phrase boundaries and accents and it is comparable with the correspondence between human labellers [7],[8]

7 Outlook

Further investigations will concern the detection of F0-Contours by statistical methods (DP-Alignment or modelling by HMM). These contours can be used to improve the classfication of accents and boundaries due to the combination of predicted accents and phrase boundaries from the text and recognition of acoustic events (like F0-movements) from the speech signal.

References

[1] A. Batliner et al.: The Prosodic Marking of Accents and Phrase Boundaries: Expectations and Results in NATA ASI: New Advances and Trends in Speech Recognition and Coding, Ed. A.J. Rubio, Bubion(Granada), June-July 993

[2] A.M. Derouault, B. Merialdo: *Natural Language Modeling for Phoneme-to-Text Transcription*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 8, pp. 742-749, Nov. 1986

[3] F. Brugnara et al.: *Automatic segmentation and labeling of speech based on Hidden Markov Models*, Speech Communication 12 (1993), 357 - 370

[4] N. Campbell: *Automatic detection of prosodic boundaries in speech*, Speech Communication 13 (1993), 343 - 354

[5] M. Lehning: Automatische Wortsegmentierung mit semikontinuierlichen Hidden Markov Modellen, Fortschritte der Akustik, DAGA 94, Dresden, Teil C, 1257 - 1260

[6] H. Ney et al.: *On Structuring Probabilistic Dependences in Stochastic Language Modelling*, Computer, Speech and Language, Vol. 8, 1994, 1 - 38

[7] M. Reyelt: *Experimental Investigation of the Perceptual Consistency and the Automatic Recognition of Prosodic Units in Spoken German*, Working Papers 41, Dept. of Linguistics and Phonetics, Lund, Sweden (1993), 238 - 241, ESCA-Workshop on Prosody, Ed. D. House and P. Touati

[8] M. Reyelt: *Consistency of Prosodic Transcriptions Labelling Experiments with Trained and Untrained Transcribers*, to appear in: Proceedings ICPHS 1995, Stockholm