DETECTION OF ACCENTS, PHRASE BOUNDARIES AND SENTENCE MODALITY IN GERMAN WITH PROSODIC FEATURES

V. Strom
e-mail: vst@asl1.ikp.uni-bonn.de

Institute of Communications Research and Phonetics (IKP),
University of Bonn,
Poppelsdorfer Allee 47, 53115 Bonn,
Germany,

ABSTRACT

In this paper detectors for accents, phrase boundaries, and sentence modality are described which derive prosodic features only from the speech signal and its fundamental frequency to support other modules of a speech understanding system in an early analysis stage, or in cases where no word hypotheses are available. A new method for interpolating and decomposing the fundamental frequency is suggested. The detectors' underlying Gaussian distribution classifiers were trained and tested with approximately 50 minutes of spontaneous speech, yielding recognition rates of 78 percent for accents, 81 percent for phrase boundaries, and 85 percent for sentence modality.

1. INTRODUCION

VERBMOBIL [9] is a multidisciplinary research project in Germany. Its goal is to develop a tool for machine translation of spoken language (the current domain is appointment scheduling) from German to English and in a later stage also from Japanese to English. The prototype will include a keyword spotting system for English and a speech understanding system for German. A prosody module (developed in Erlangen and Munich, [2][3]) that gets its information from the acoustic signal and the word hypothesis generator is part of the speech understanding component.

VERBMOBIL also investigates an innovative and highly interactive architecture model for speech understanding. One of its guidelines is incrementality, e.g. every module should process its input with minimum delay. For this architecture an experimental system was designed that also has a prosody module. Currently this module uses only the speech signal and its fundamental frequency as input.

Not using the output of the word hypothesis generator means that no normalized duration features can be obtained, since the intrinsic syllable duration can only be determined when the spoken words are known. A detector using duration features should perform better than one which does not, but it cannot be applied in cases where no word hypotheses are available.

In the BELLEx3+1 system [1] the accent detector described here informs the morphologic-prosodic parser, a part of the word recognizer BELLEx3+1. Thus, it cannot use word hypotheses.

The VERBMOBIL prototype will only roughly follow the *English* part of a dialogue: The dialogue manager classifies utterances into speech acts like DATE SUGGESTION or REJECTION using just the output of the key word spotter. The sentence modality detector described here can be used to segment utterances consisting of more than one sentence.

2. MATERIAL

A subset of spontaneous spoken dialogues collected for the VERBMOBIL project has been prosodically labelled on three levels: the functional level and the 1st and 2nd perceptive levels [7]. On the functional level sentence modality and accents are labelled. On the first perceptive level the prosodic structuring is labelled. Full prosodic phrases are distinguished from intermediate phrases. The second perceptive level describes intonation: Every accent and phrase boundary gets a tone label very similar to those in the ToBI system [8].

An automatic phoneme segmentation was used to obtain the time alignment of vowel and syllable boundaries. The fundamental frequency was determined with the `get_f0` program of ESPS[1]. Neither segmentation errors nor $F0$ errors were corrected manually, but turns containing automatically detectable segmentation errors were discarded. The remaining data base contains approximately 50 minutes of speech.

3. FEATURE EXTRACTION

To get a parametrization of the fundamental frequency and energy contours suitable for direct classification in an incremental way, for every 10 ms frame eleven features are calculated that describe the fundamental frequency and energy contours in a certain neighbourhood.

3.1. $F0$ interpolation

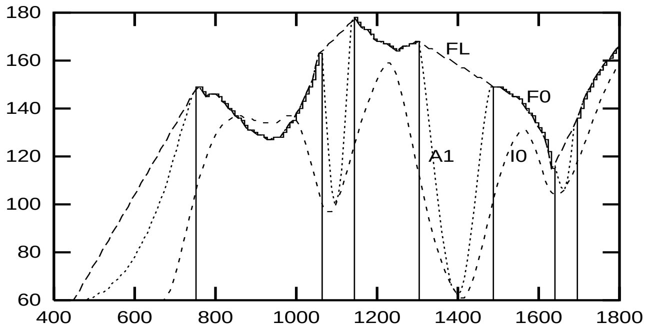

Figure 1. *The solid line is the $F0$ in Hz,* **I0** *the low pass filtered $F0$,* **FL** *the linear interpolated $F0$, and* **A1** *is the* **I0** *"adapted" to* **FL**; *explanation in the text.*

For unvoiced segments, $F0$ was defined to be zero. The interpolation of the fundamental frequency $F0$ works as follows: The initial steps are the low pass filtering of the $F0$, and the linear interpolation of the $F0$ (with a ramp function from/to the zero line at the first and last voiced

[1]Entropic Signal Processing System

region of an utterance). This low pass filtered $F0$ is reset to the original values in voiced regions and adapted to the linear interpolated $F0$ in unvoiced regions using a hanning window to force continuity. Figure 1 shows the effect of this adaption.

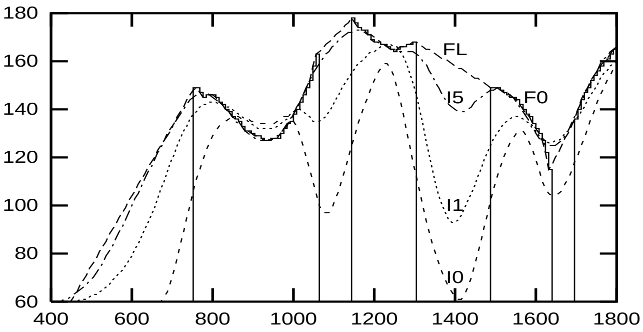

Figure 2. *The solid line is the $F0$ in Hz,* **I0** *the low pass filtered $F0$ and* **FL** *the linear interpolated $F0$.* **I1** *and* **I5** *are the output of the first and the fifth iteration step;* **I1** *also is the low pass filtered* **A1** *from figure 1.*

This adapted $F0$ is low pass filtered again, reset in voiced regions and adapted in unvoiced regions. Figure 2 shows the initial curves and the output of the first and fifth iteration steps. This algorithm still works incremental, its delay, however, is the sum of the filter dalays.

3.2. $F0$ decomposition

The interpolated $F0$ was then decomposed into two components using band-pass filters, the one component describing the global behaviour of the $F0$ contour and the other one the local behaviour, by analogy with the phrase and accent components in [4]. Now three components are filtered out with edge frequencies optimized with respect to the accent recognition rate.

The interpolated $F0$, its three components, and the time derivatives of those four functions (calculated by a regression line over 200 ms) yield eight $F0$ features.

3.3. Energy features

Furthermore three energy features are calculated that were used for syllable nucleus detection in [5]: the so-called nasal band (50-300 Hz), the sonorant band (300-2300 Hz), and the fricative band (2300-6000 Hz). These features are obtained by short time FFT followed by median smoothing.

4. DETECTION OF ACCENTS

For every frame one of five classes was derived from the prosodic labels and the automatic phoneme segmentation: 1) Not a vowel, 2) a vowel within an unaccented syllable, 3) a vowel within an accented syllable labelled with phrase accent, 4) within a secondary accent or 5) emphasis. A Gaussian distribution classifier with a special cost function was used to map each frame on one of the superclasses "accented vowel yes/no" (**A/NA**), followed by a filter that suppresses "accented regions" shorter than six frames.

figure 3 shows the accent detector output for an utterance together with all labels, the interpolated $F0$, their components and the energy features.

The classifier yields one accent label every frame, the evaluation, however, was carried out syllable by syllable: If within an accented syllable at least one frame got an **A**-label, or within an unaccented syllable not a single frame the **NA**-label, this syllable was considered to be correctly classified. This means that the detector did not need to detect the exact vowel positions. The results are shown in table 1.

	classified as		r.f.
	A	NA	
A	66.53	33.47	35.39
NA	23.45	76.55	74.61

Table 1. *Confusion matrix of the detector. Recognition rates is 74.01 percent, average recognition rate is 71.54 percent, with 2 classes and 9887 syllables. r.f is the (relative) frequency of occurance; all figures are in percent.*

5. DETECTION OF PHRASE BOUNDARIES AND SENTENCE MODALITY

The original labels disinguish between full prosodic phrase boundaries **B3**, intermediate phrase boundaries **B2**, nongrammatical phrases boundaries **B9**, and every other word boundary **B0**. As words are not known to the detector described here, **B0** is used for "every other syllable boundary". During the analysis syllable boundaries are obtained by the syllable nucleus detector mentioned above.

Every **B3** boundary may have a functional label for "question" **Q**. If not, it is interpreted as end of "statement" **S**, if the boundary tone is low (L%), or "progredient" **P** otherwise.

The detector views a window of (if possible) four syllables, following [3]. The detector output refers to the syllable boundary between the second and the third syllable nucleus in the case of a 4-syllable window. For each window a large feature vector is constructed: The 11 features (as described in section 3) at each of the 4 syllable nuclei and 7 time features (the lengths of the four syllable nuclei and the distances between them) make 53 features. The 30 best of them have previously been determined with a feature selection algorithm as described in [6].

A Gaussian distribution classifier was trained to distinguish between all combinations of boundary types and tones. The classifier output was then mapped on the boundary types **B0, B2, B3** or **B9**, and on the sentence modality **Q, S** or **P** if not a **B3** was detected, and on **0** otherwise. Thus, both detectors use the same underlying classifier.

At the beginning of an utterance a 3-syllable window is applied, at the end of an utterance (which is defined here as "no output from the syllable nucleus detector for 500 ms") a 3-syllable and a 2-syllable window are applied. Two more window types are necessary for utterances consisting of only one or two syllables. For each of the 6 window types a seperate classifier is used.

	classified as				rel.freq.
	B0	B2	B3	B9	
B0	87.48	2.18	8.41	1.93	76.83
B2	56.80	32.93	7.25	3.02	4.16
B3	26.07	1.90	67.35	4.68	15.87
B9	34.80	1.20	16.40	47.60	3.14

Table 2. *Confusion matrix of the classifier for phrase boundary detection. Recognition rate is 80.76 percent, average recognition rate is 58.84, with 4 classes and 9887 syllables.*

Strictly speaking the detector output does not refer to the syllable boundary but to the time interval between the second and the third *detected* syllable nucleus (in the case of a 4-syllable window). If a **B2,B3** or **B9** was within this interval, the large feature vector used for training and

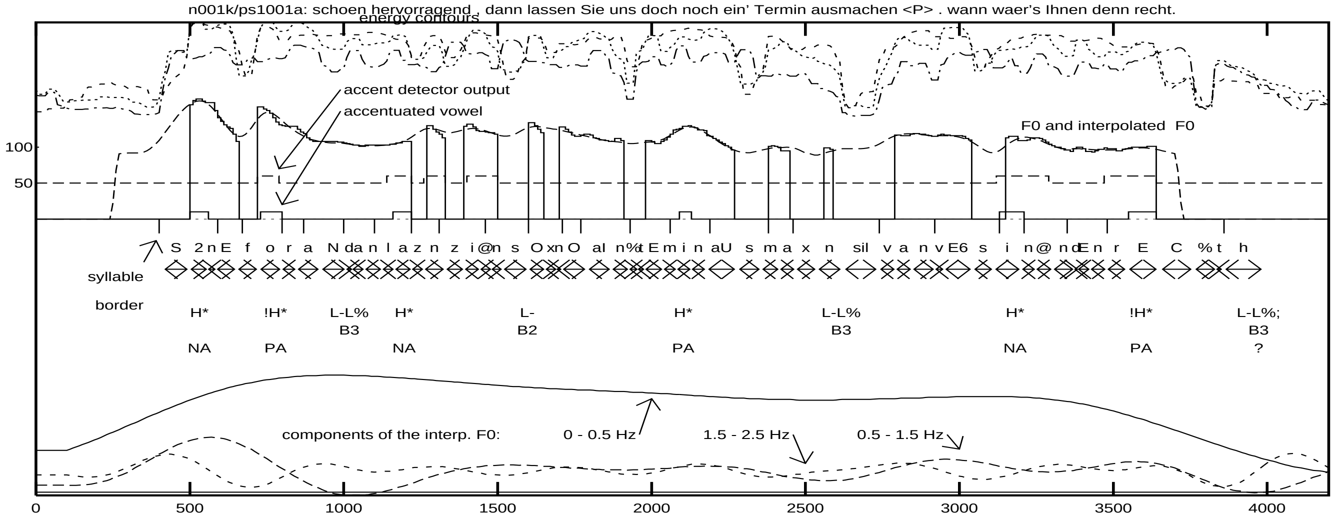

Figure 3. *Feature extraction and accent detection; the time axis is scaled in ms. From top to bottom: The three energy measures, the F0 and the interpolated F0 (the y-axis is scaled in Hz and referes only to these two curves), the accent detector output and the "accentuated vowel yes/no" label as square functions, the phoneme labels (SAMPA), the ToBI labels, the phrase structuring, the functional labels (PA denote the primary accents, NA the secondary accents), and the three components of the interpolated F0.*

	classified as				rel.freq.
	0	P	S	Q	
0	91.35	1.85	6.16	0.64	84.13
P	53.01	30.87	7.65	8.47	4.60
S	30.23	1.27	65.40	3.11	8.90
Q	2.13	5.32	32.98	59.57	2.36

Table 3. *Confusion matrix of the classifier for sentence modality detection. Recognition rates is 85.50 percent, average recognition rate is 61.90 percent, with 4 classes and 9887 syllables.*

testing the classifier got the appropriate label, else it got the **B0** label. The results are shown in tables 2 and 3.

6. CONCLUDING REMARKS

The recognition rates of the prosody detectors as stand alone modules seem good enough to contribute posetively to the overall performance, but this has still to be proved.

Currently an improved version of the syllable nucleus detector is under work. It will allow us to make the phrase boundary detection dependant from the syllable nucleus detection, and thus to use the intonation labels for training and a wider context for feature extraction.

ACKNOWLEDGEMENTS

The author gratefully acknowledge the disposition of classification software from H. Niemann, University of Erlangen–Nürnberg.

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01 IV 101 G. The responsibility for the contents of this study lies with the author.

REFERENCES

[1] F. Althoff, J. Carson-Berndsen, G. Drexel, D. Gibbon, K. Hübener, U. Jost, K. Kirchhoff, M. Pampel, A. Petzold, and V. Strom BELLEx3+1 Linguistische Worterkennung unter Berüksichtigung der Prosodie internal Verbmobil Technisches Dokument Nr. 22, Universität Bielefeld, Universität Bonn, Universität Hamburg, 1995.

[2] A. Kießling, R. Kompe, A. Batliner, H. Niemann, and E. Nöth. Automatic labeling of phrase accents in german. In *Proc. Int. Conf. on Spoken Language Processing*, pages 115–118, Yokohama, 1994.

[3] R. Kompe, A. Batliner, A. Kießling, U. Kilian, H. Niemann, E. Nöth, and P. Regel-Brietzmann. Automatic classification of prosodically marked phrase boundaries in German. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 173–176, Adelaide, 1994. 115–118, Yokohama, 1994.

[4] H. Mixdorf and H. Fujisaki. Analysis of voice fundamental frequency contours of german utterances using a quantitative model. In *Proc. Int. Conf. on Spoken Language Processing*, pages 2231–2234, Yokohama, 1994.

[5] E. Nöth. *Prosodische Information in der automatischen Spracherkennung.* Max Niemeyer Verlag, Tübingen, 1991.

[6] H. Niemann. *Klassifikation von Mustern.* Springer Verlag, Berlin, 1983.

[7] M. Reyelt. and A. Batliner Ein Inventar prosodischer Etiketten für VERBMOBIL. internal Verbmobil Memo Nr. 33, TU Braunschweig, Ludwig-Maximilian-Universität München, 1994.

[8] K. Silverman, M. Beckman, J. Pitrelli, M. Osterndorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. Tobi: A standard for labeling english prosody. In *Proc. Int. Conf. on Spoken Language Processing*, pages 867–870, 1992.

[9] W. Wahlster. Verbmobil – Translation of Face-To-Face Dialogues. In *Proc. European Conf. on Speech Communication and Technology*, volume "Opening and Plenary Sessions", pages 29–38, Berlin, 1993