WHAT'S IN THE "PURE" PROSODY?

V. Strom and C. Widera
e-mail: vst@ikp.uni-bonn.de

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

ABSTRACT

Detectors for accents and phrase boundaries have been developed which derive prosodic features 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. The detectors' underlying Gaussian distribution classifiers were trained with 50 minutes and tested with 30 minutes of spontaneous speech, yielding recognition rates of 74% for accents and 86% for phrase boundaries. Since this material was prosodically hand labelled, the question was, which labels for phrase boundaries and accentuation were only guided by syntactic or semantic knowledge, and which ones are really prosodically marked. Therefore a small test subset has been resynthesized in such a way that comprehensibility was lost, but the prosodic characteristics were kept. This subset has been re-labelled by 11 listeners with nearly the same accuracy as the detectors.

1. INTRODUCION

VERBMOBIL [16] 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, [5][4]) 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. For this architecture an experimental system was designed that also has a prosody module. This module uses only the speech signal and its fundamental frequency as input. The accent detector in this module can not use word hypotheses since it is *part* of the word recognizer [1].

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. A phrase boundary detector that needs no word hypotheses can be used to segment utterances consisting of more than one speech act.

Prosody recognition without word hypothesis means that no normalized duration features can be obtained, since the intrinsic syllable duration can only be determined when the spoken words are known.

The question then was, which of the labelled accents and phrase boundaries can be recognized by human or machine, if no word information is available.

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 [2]. On the functional level[1] sentence modality and accents are labelled. "Primary accents" (**PA**) were distinguished from "secondary accents" (**NA**) and "emphasis" (**EK**) according to their prominence, but this distinction is not made in this study.

On the first perceptive level the prosodic structuring is labelled. Full prosodic phrases (**B3** boundaries) are distinguished from intermediate phrases (**B2** boundaries). Furthermore irregular phrase boundaries are labelled with **B9**.

B3 boundaries are associated with a clear F0 reset and possibly a short pause, **B2** boundaries with a less clear F0 reset. **B9**s are labelled at sentence interruptions, hesitations, etc. They are often associated with a longer pause or a filled pause. This study deals only with the **B3** boundaries.

The second perceptive level describes intonation: every accent and phrase boundary gets a tone label very similar to those used in the ToBI system [14]. These labels were used as explicit clustering during training the detectors.

The procedure for labelling was as follows: First, the whole utterance was listened to and the **B3** labels were set. Afterwards the phrases were labelled *separately*, first the accents, then the intonation, the intermediate and irregular phrase boundaries.

An automatic phoneme segmentation was used to obtain the time alignment of vowels and syllable boundaries. The fundamental frequency was determined with the `get_f0` program of ESPS[2].

3. ACCENT DETECTION

To obtain a parameterization of the fundamental frequency and energy contours suitable for direct classification, eleven features are calculated per frame that describe the fundamental frequency and energy contours in that region.

First, $F0$ is interpolated in unvoiced segments by an iterative method based on low pass filters and linear interpolation to obtain a steady, smooth contour. Then it is decomposed by band pass filters. The components describe the $F0$ contour globally and locally. The interpolated $F0$, its three components, and the time derivatives of those four functions yield eight $F0$ features.

Furthermore three energy features are calculated that were used for syllable nucleus detection in [10]: the so-called nasal band (50-300 Hz), the sonorant band (300-2300 Hz), and the fricative band (2300-6000 Hz). These

[1] Of course labels on this level are also based on perception. Therefore the accent labels are not alway identical to the semantic focus.

[2] Entropic Signal Processing System

features are obtained by short-time FFT followed by median smoothing.

A Gaussian distribution classifier with a special cost function was trained to destinguish between "accented vowel yes/no" (**A/NA**) every 10 ms frame, followed by a filter that suppresses "accented regions" shorter than six frames. The evaluation 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 **A**-label, this syllable was considered to be correctly classified.

In the test set 2691 of the 10601 syllables were accented. 1790 have been detected, 1854 inserted, which corresponds to a recognition rate of 74.00% or an accuracy of –2.38% ((1790–1854)/2691).

4. PHRASE BOUNDARY DETECTION

The phrase boundary detector views a window of (if possible) four syllables. Its output refers to the syllable boundary between the second and the third syllable nucleus (in the case of a 4-syllable window). Syllables are found by a syllabic nucleus detector based on energy features derived from the speech signal.

For each window a large feature vector is constructed: The 11 features as described in the previous section at each of the 4 syllable nuclei in the window, plus 7 time features (the lengths of the four syllable nuclei and the distances between them). The 30 best features have previously been determined with a feature selection algorithm as described in [11].

A Gaussian distribution classifier was trained to distinguish between all combinations of boundary types and tones. The classifier output was then mapped on the decision "**B3** yes/no".

In the test set 1959 of 10601 syllable boundaries were B3s. 1139 were correctly detected, 314 inserted. This corresponds to a recognition rate of 86.29%, or an accuracy of 42.11% ((1139–314)/1959). If the 251 turn-final **B3**s are not counted, the accuracy is 33.6%.

5. DELEXICALIZATION

The prosody module developed in Erlangen and Munich achieved better recognition results on the same data [6] because it had the word hypotheses graph as additional input. It uses the normalized duration as a feature (instead of the duration in ms as in the detector described above) and additionally applies a language model based on word categories and prosodic labels.

Since that language model involves syntactic and semantic knowledge, the question was, can accents and boundaries marked by pure prosodic means be detected without word information. Or, in other words, wether the results reported above still could be improved with the approach described in section 3 and 4.

One approach is to delexicalize the speech, that means to resynthesize it in such a way that comprehensibility gets lost, but the prosodic characteristics are kept, and then to re-label it.

Different ways of delexicalization have been proposed. Spectral inversion [9] (the sign of every second sample is inverted), rigid band pass filtering [8][7][15], and LPC resynthesis after setting the formants to neutral schwa-like values [3].

We tried two methods that will be described in the next two sections.

6. SAWTOOTH SIGNALS

Both spectrally inverted and band pass filtered signals still contain segmental information. The problem with LPC based techniques is the automatic formant tracking. Either the result is poor, or time consuming manual correction is required.

Therefore we decided first to replace voiced segments by a sawtooth signal of the same pitch and energy, and unvoiced segments by silence. Sawtooth signals sound relatively human-like, and a pitch marker good enough for our purpose was available [13].

The problem with the methods of delexicalization described so far is that they make re-labelling difficult since they all destroy segment boundaries (except voiced/unvoiced boundaries). Boundaries between syllable- and word-like segments are lost. Furthermore we did not want to present visual information such as $F0$ contour since they might influence listeners' judgements.

6.1. Labelling procedure

We decided to label auditorily by key stroke. Phrase boundaries and accents were labelled separately to imitate the procedure of the original labelling (see section 2).

11 listeners were asked to strike a key immediately if they perceive the first phrase boundary. The term phrase boundary was not explained in detail. The utterance was cut at that point. The listener could replay the first part and the rest of the utterance, move the cut forward or backward, or repeat cutting. After confirming that label, she/he proceeded in the same way with the rest of the utterance. Therefore reaction time did not influence the results. If a **B3** was within an unvoiced segment, a label within the same unvoiced segment was counted as correct. Otherwise it has to be not further than 50 ms from the **B3**.

Accents were also labelled by key stroke. For auditory checking, a short beep was superimposed at the appropriate point, and again the listener could repeat the labelling or move the label before confirming it. Labels were counted as correct if they were within a syllable carrying a **PA** or **NA** label (**EK** did not occur in the subset).

The 11 listeners were staff members and some of them had experience in prosodic labelling. Both accent and phrase boundary labelling was divided into two sessions, preceded by an instruction and training phase. Each session took approximately 45 minutes.

6.2. Phrase boundaries

For phrase boundary re-labelling a subset of 20 utterances was selected containing 58 **B3** boundaries. The utterance final **B3**s were not taken into account. 16 **B3**s had an L-L% label, 11 an H-H%, 7 an H-L%, and 4 an L-H%. 12 further boundaries were labelled as **B2**, and 8 as **B9**.

The detector recognized 21 of the 38 non-final **B3**s and inserted 11 (accuracy 26%). On average the 11 listeners perceived 23 of the **B3**s as phrase boundaries and inserted 10.3 boundaries where neither a **B3**, a **B9** nor a **B2** had been labelled (mean accuracy 33%). Insertion errors of the listeners were not counted as strictly as those of the detector because no definition of the term phrase boundary was given to the listeners.

All **B3**s associated with a pause were found by the listeners, and nearly all of them by the detector. Other characteristics like phrase and boundary tone do not seem to play a role.

Since the prosodic realization of a **B9** is often similar to that of a **B3**, in a further evaluation both boundary types were treated equal: Now the listeners recognized 26 of the 46 **B3**s and **B9**s on average, but the accuracy rose only slightly to 36%.

The detector found 2 of the 8 additional boundaries, but made 2 more insertion errors leading to a slightly lower accuracy of 22%.

6.3. Accents

For accent re-labelling a subset of 22 phrases was selected containing 51 accented syllables, 24 **PA**s and 27 **NA**s. The detector recognized 33 of them and inserted 18 (accuracy 29%). On average the 11 listeners perceived 27.6 accents and inserted 16.6 (accuracy 21%).

PAs are both perceived and recognized more reliably than **NA**s, but again it seems to be irrelevant if the accent was low or high, or where the peak lay within the syllable.

7. NONSENSE SPEECH

Since labelling by key stroke is quite hard for non-drummers (accent labelling of 22 utterances and phrase boundary labelling of 20 utterances took approximately 30 minutes each), we were concerned that this might have influenced the results.

Therefore we looked for a method of delexicalization that preserves segment boundaries. This allowed transcription of the delexicalized utterances and relabelling with paper and pencil.

We used a PSOLA-like synthesizer [12] to produce nonsense sentences. The synthesizers input was a phoneme string with duration and $F0$ values. The phoneme string was obtained from the automatic phoneme segtmentation: every phoneme was replaced by one of the same cathegory randomly (i.e. vowels by vowels, voiced plosives by voiced plosives, etc.), but with respect to German phonotactics. We did not manipulate the unit's amplitude because the resolution of the automatic phoneme segmentation (10 ms) was not accurate enough to obtain reasonable energy values.

7.1. Phrase boundaries

The 20 utterances described in section 6.1 were delexicalized in this manner, and a German transcription of these nonsense utterances was presented to the 11 listeners (syllables were separated by a blank). They could listen to each nonsense utterance as often as they wanted. The evaluation criteria were the same as in section 6.1.

On average 23.7 of the 38 **B3**s were perceived, this is nearly the same figure as in section 6.1, but only 6.8 **B3**s were inserted. This leads to a significantly higher accuracy than with the sawtooth signals (see figure 1). Including the **B9**s in the evaluation did not improve the accuracy.

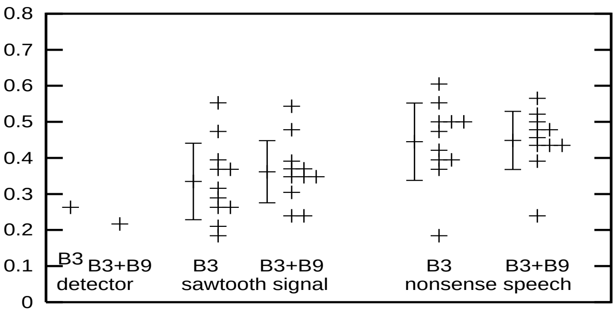

Figure 1. *Accuracy of the phrase boundary detector and the listeners' phrase boundary labels. Mean and standard deviation are illustrated as errorbars.*

7.2. Accents

Delexicalisation, transcription and re-labelling of the 22 phrases was carried out as in section 6.3. On average 26.9 of the 51 accents were perceived and 15.2 inserted. The mean accuracy was 23% (see figure 2).

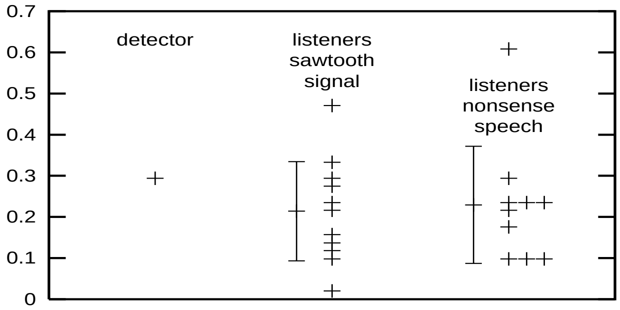

Figure 2. *Accuracy of the accent detector and the listeners' accent labels. Mean and standard deviation are illustrated as errorbars.*

8. DISCUSSION

Detectors for accents and phrase boundaries have been described which do not use word information but "pure" prosodic features: a paramerterized description of F0 and energy contour and unnormalized time features. The question was, can these detectors be improved. Therefore delexicalized utterances, which contained only the pure prosody, were presented to labellers.

The method of delexicalisation and labelling is not crucial, with one exception: in nonsense speech the listeners inserted less boundaries. This suggests that phonotactic knowledge makes perception of phrase boundaries easier.

In delexicalized speech, the detector is nearly as good as humans in prosodic labelling, slightly better at accent labelling, and slightly worse at phrase boundary labelling. We cannot explain why the accent detector is better than the average listener. The phrase boundary detector probably does worse because it uses a context of only four syllables and it does not expect boundaries to occur after certain time intervals as humans do.

Perception of accents and phrase boundaries without understanding is surprisingly difficult. Therefore we believe that detection with pure prosodic features cannot be substantially improved.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the use of classification software from H. Niemann, University of Erlangen–Nürnberg, and of the pitch period marker from Ansgar Rinscheid, University of Bochum.

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 D 08. 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. Linguistische Worterkennung unter Berücksichtigung der Prosodie. Verbmobil Technisches Dokument Nr. 22, Universität Bielefeld, Universität Bonn, Universität Hamburg, 1995.

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

[3] J.R. de Pijper and A.A. Sandermann. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. In *J. of the Acoustic Society of America*, volume 96, pages 2037–2047, 1994.

[4] R. Kompe, A. Batliner, U. Kilian, H. Niemann, E. Nöth, and P. Regel-Brietzmann. Automatic labeling of prosodically marked phrase boundaries in German. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 173–176, Adelaide, 1994.

[5] 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.

[6] R. Kompe. Prosodic scorung of word hypotheses graphs. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1333–1336, Madrid, 1995.

[7] J. Kreimann. Perception sentence and paragraph boundaries in natural conversation. In *Journal of Phonetics*, volume 10, pages 163–175, 1982.

[8] I. Lehiste. Perception of sentence and paragraph boundaries. In B. Lindblom and S. Ömann, editors, *Frontiers of speech communication research*, pages 191–201. Academic Press, New York, 1979.

[9] I. Lehiste and W. S-Y Wang. Perception of sentence boundaries with and without semantic information. In W. Dressler and O. Pfeiffer, editors, *Phonologica*, volume 19, pages 277–283. Innsbruck, 1976.

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

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

[12] T. Portele, B. Steffan, R. Preuss, W.F. Sendelmaier, and W. Hess. Hadifix - a speech synthesis system for German. In *Proc. Int. Conf. on Spoken Language Processing*, pages 1227–1230, 1992.

[13] A. Rinscheid. Automatische bestimmung von periodenmarken mit dem emark-Algorithmus. In *Proc. Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik – DAGA*, pages 1048–1051, Frankfurt a. M., 1993. DPG–GmbH.

[14] 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.

[15] D. Schaffer. The role of intonation as a cue to topic management in conversation. In *Journal of Phonetics*, volume 12, pages 327–344, 1984.

[16] W. Wahlster. Verbmobil – translation of face-to-face dialogs. In *Proc. European Conf. on Speech Communication and Technology*, volume Opening and Plenary Sessions, pages 29–38, Berlin, 1993.