UNSUPERVISED EXTRACTION OF PROSODIC STRUCTURE

Uwe D. Reichel

Research Institute for Linguistics, Hungarian Academy of Sciences
uwe.reichel@nytud.mta.hu

Abstract: Our approach for unsupervised extraction of prosodic structure in spontaneous speech consists of the four steps: chunking into interpausal units, syllable nucleus extraction, prosodic boundary detection, and pitch accent detection. The extraction is based on acoustic features derived from F0 parameterization, and on energy and segment duration features. Phrase boundaries and accents are detected by means of nearest centroid classifiers which are bootstrapped from the data.

1 Introduction

Manual prosodic annotation of prosodic boundaries and pitch accents is a time consuming task. It requires inter- and intra-labeler consistency checks [1, 2], and since it is often embedded in a particular theoretic framework it often cannot straight-forwardly be adopted to new language data. Thus an automatization of the annotation process is highly desirable. So far most studies addressed the prediction of prosodic structure from linguistic and/or acoustic features by supervised learning, e.g. [3, 4, 5, 6, 7]. Considerably speeding up the annotation process, the bottleneck of these approaches is still their need for manually annotated training data, which is often not available especially for under-studied languages.

So far much less studies focused on the *unsupervised* extraction of prosodic structure. To our knowledge, at the current state the best results were obtained by [8] who applied a continuous wavelet transform to a composite signal combining F0, energy, and word duration. Pitch accents are identified by a high amount of co-occurring maxima of wavelets on different scales, and analogously phrase boundaries by co-occurring minima. [9] propose an iterative approach: after an initial clustering based on acoustic features, conditional probability distributions over linguistic features are derived from reliable items close to their respective centroid. The probabilities are then used to classify items with less acoustic evidence to belong to one of the prosodic classes. [10] propose an iterative parallel training and application of weak classifiers on acoustic and linguistic features, that starts with a small amount of manual seed annotations and iteratively automatically annotates the rest of the data.

The current study aims to contribute to this line of research by an approach that relies on acoustic features only and does not require manually annotated seed exemplars but initializes clustering in a purely data-driven bootstrap approach. It is applied to spontaneous speech instead of read news speech used in the unsupervised extraction studies mentioned above.

2 Data

2.1 Corpus

The underlying data consists of the prosodically annotated German parts of the Verbmobil I corpus [11] of spontaneous dialog speech, that is available in the BAS repository [12]. The used part comprises 180 turns (22 minutes, 10 speakers). The annotation relevant for this study

contains an automatic signal-text alignment on the phoneme and word level by MAuS [13] (tier *MAU*) and a manual prosodic annotation (tier *PRB*). From the compound prosodic labels the prosodic event strength encodings were translated into 2 classes each for boundaries and accents: presence (class 1) vs. absence (class 0) of a prosodic event. For prosodic boundary detection each word boundary was marked accordingly, mapping no label and B2 (minor boundary) cases to class 0, and B3 and B9 (major, irregular boundary) cases to class 1. For accent detection no label corresponds to class 0, and NA, PA, EK (weak, strong, emphatic accent) were assigned to class 1.

2.2 F0 and energy extraction

F0 was extracted by autocorrelation (PRAAT 5.3.16 [14], sample rate 100 Hz). Voiceless utterance parts and F0 outliers were bridged by linear interpolation. The contour was then smoothed by Savitzky-Golay filtering [15] using third order polynomials in 5 sample windows and transformed to semitones relative to a base value. This base value was set to the F0 median below the 5th percentile of an utterance and serves to normalize F0 with respect to its overall level.

Energy in terms of root mean squared deviation was calculated with the same sample rate as F0 in Hamming windows of 50 ms length.

3 Tasks and general procedure

The tasks of this study are: chunking of the utterances into interpausal units, syllable nucleus assignment, prosodic phrase boundary detection, and pitch accent detection.

2-class (presence or absence) phrase boundary decisions are made on the word level for each word boundary, and 2-class accent decisions on the syllable level for each syllable nucleus. Since each step operates partially on the output of the previous steps, the tasks are carried out sequentially.

In an initial exploratory phase a third of the data (60 turns; in the following referred to as *tuning data*) was used to successively tune some of the syllabification, phrasing, and accent localization parameters in a brute-force way testing several parameter values. This exploratory phase was needed to gain first experience how to initialize the components.

4 Chunking and syllable nucleus extraction

4.1 Chunking

In the current approach the chunking of a speech signal into interpausal units is derived simply from inverting the output of a pause detector described in more detail in [16].

Pauses were detected by energy (RMS) comparison of the low-pass filtered signal between an analysis window w_a and a longer reference window w_r with the same time midpoint that are moved along the signal in 50ms steps. Low-pass filtering was carried out by a Butterworth filter of order 5. For pause assignment the energy in w_a has to be lower than in w_r by a factor v, i.e. $\mathrm{RMS}(w_a) < \mathrm{RMS}(w_r) \cdot v$. The model parameters are: the upper cutoff frequency f, the window lengths w_a, w_r, the threshold factor v, and the minimum required pause length l in order not to erroneously consider the occlusion phase of plosives as pauses and to extract pauses long enough to justify separating the signal into different chunks. The parameters were estimated in a previous study [16] by the non-linear Nelder-Mead Simplex optimization [17] that yielded the following values: $f = 8000$Hz, $w_a = 0.15s$, $w_r = 5s$, $v = 0.08$, $l = 0.5$.

4.2 Syllable nucleus extraction

Syllable nucleus assignment follows to a large extent the procedure introduced in [18]. Again an analysis window w_a and a reference window w_r with the same time midpoint were moved along the this time band-pass filtered signal in 50ms steps. Filtering again was carried out by a 5th order Butterworth filter with the cutoff frequencies 200 and 4000Hz. For a syllable nucleus assignment the energy in the relevant frequency range r is required to be higher in w_a than in w_r by a factor v, and additionally had to surpass a threshold x relative to the maximum energy RMS_{max} of the utterance, i.e. $\mathrm{RMS}(w_a) > \mathrm{RMS}(w_r) \cdot v \wedge \mathrm{RMS}(w_a) > \mathrm{RMS}_{max} \cdot x$. Exploratory tuning yielded the following values: $w_a = 0.05s$, $w_r = 0.11s$, $v = 1.1$, $x = 0.1$.

5 Prosodic structure assignment

After chunking and syllable nucleus assignment prosodic structure was induced in terms of bootstrapped nearest centroid classification. The feature sets, feature weighting, and the chunking procedure are described in the following.

5.1 Phrase boundary features

The feature set for phrase boundaries was derived from a parameterization of pitch register discontinuity at each word boundary as illustrated in Figure 1. Within a stylization window of maximally 4 seconds centered on a word boundary and limited by the chunk boundaries three regression lines, a base- a mid- and a topline were fitted to three F0 contour segments: to the segments left- and right adjacent to the boundary, seg_1 and seg_2, and to their concatenation seg_{12}. The midline represents the F0 register level, and a linear regression through the pointwise distances between base and topline represents the register range. Discontinuity is then defined in terms of the deviation of the level and the range regression lines between seg_{12} and seg_1 and seg_2, respectively. This approach is described in more detail in [19]. From this parameterization the following discontinuity features were extracted:

- the pairwise absolute slope differences of the level and range regression lines between the 3 segments: s_{1_12}, s_{2_12}, s_{1_2},
- the RMS between these lines,
- the absolute pitch resets d_{1_12}, d_{2_12}, d_{1_2}.

In [19] these features had turned out to be positively correlated with perceived prosodic boundary strength.

Next to these register discontinuity features final lengthening was captured by the normalized duration of the last vowel in the *MAU* tier preceding the word boundary. Normalization consisted in dividing the length of this vowel by the mean length of all vowel segments with the same label in the entire data.

5.2 Accent features

Accent features were extracted for each detected syllable nucleus. Within an analysis window of 0.2s length centered on the nucleus the F0 maximum, median, inter quartile range, and RMS were calculated. Analogously, for the energy contour within this window we calculated the maximum, median, and the RMS. All values were normalized within a longer normalization window of maximum length 0.6s with the same center and limited by the boundaries of the underlying extracted prosodic phrase.

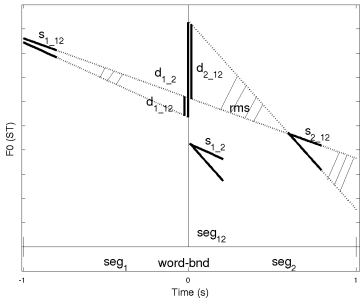

Figure 1 – Word boundary parameterization by pitch discontinuity features that consist in slope s_*, RMS, and reset d_* deviations between the F0 registers left- and right-adjacent to the word boundary (seg_1, seg_2) as well as their deviations from a common trend.

Next to these standard features it was quantified, to what extent the local register measured around the syllable nucleus sticks out of the corresponding register stretch of the prosodic phrase. This Gestalt property was modeled as illustrated in Figure 2. As described in section 5.1 for the prosodic phrase, also within the analysis window around the pitch accent candidate, a base-, a mid-, and a topline were fitted, as well as a range regression line for the base-topline pair. Then the deviations of the locally fitted mid- (i.e. level) and range line from the corresponding portions of prosodic phrase-level lines were measured in terms of RMS. The higher the RMS, the more the local register around the pitch accent sticks out of the general register trend within the prosodic phrase. This parameterization was used e.g. in [20] in order to characterize accentual phrases.

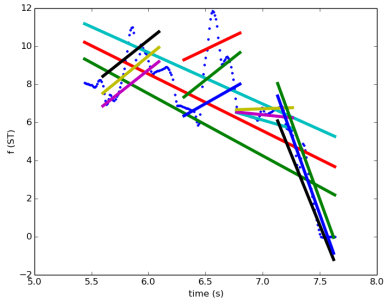

Figure 2 – Gestalt parameterization: deviation of the local register level and range around the pitch accent candidate from the underlying prosodic phrase.

Finally, adopting the superpositional framework of [16], which is illustrated in Figure 3 the local F0 shape was parameterized in the analysis window by means of a third order polynomial after normalizing it pointwise to the F0 range of the corresponding portion in the prosodic phrase. Range normalization to some extent abstracts from the influence of F0 declination, so that pitch accent candidates are comparable across different positions within a prosodic phrase. The absolute polynomial coefficient values were added to the feature pool.

In addition to the F0 and energy features, the normalized duration of the syllable nucleus vowel was added to the pool as described in section 5.1.

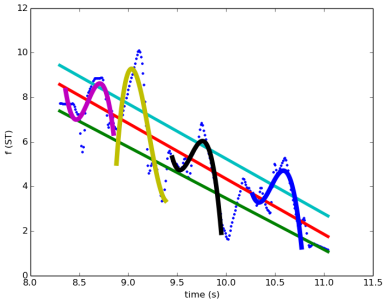

Figure 3 – Superpositional F0 representation of global and local components. The local F0 shape around a syllable nucleus is parameterized by a third order polynomial. The polynomial is fitted to range-normalized F0 values, that is, F0 is pointwise normalized relative to the values [0, 1] defined by the corresponding local base- and topline point.

5.3 Clustering and feature weighting

Similar to [10, 21] the prosodic event detection is carried out by means of clustering. However, instead of initializing the clustering by hand-annotated data as in [10], the cluster centers are bootstrapped from the data based on few assumptions. For boundary detection these assumptions are: (1) each pause is preceded by a boundary, and (2) since prosodic phrases have a minimum length, in the vicinity of pauses there are no further boundaries. Thus cluster centroids were initialized by assigning class 1 (prosodic boundary) to the feature vectors at word boundaries left-adjacent to a pause, and class 0 (no boundary) to all other word boundaries within 1 second preceding and following a pause.

For accent assignment the two assumptions are: (1) all words longer than a threshold t_a are likely to be content words that contain a high amount of information and are thus taken as class 1 (accented) representatives, and (2) all words shorter than a threshold t_{na} are likely to be function words with a low amount of lexical information and are thus taken as class 0 (no accent) representatives. In the current study t_a and t_{na} were set to 0.6s and 0.15s, respectively. For words fulfilling criterion (1) the most prominent syllable (approximated by the sum of its feature values) was added to the class 1 cluster. For words fulfilling criterion (2) all syllables were added to the class 0 cluster.

From this initial clustering feature weights were calculated from the mean cluster silhouette derived separately for each feature. The weights thus reflect how well a feature separates the seed clusters.

After this cluster initialization the remaining items (word boundaries for phrase boundaries, and syllable nuclei for pitch accents) are assigned to the classes 0 or 1 in a single pass the following way: for each feature vector i its weighted Euclidean distances $d_{i,0}$ and $d_{i,1}$ to the class 0 and class 1 centroids are calculated, and the quotient of both distances $q_i = \frac{d_{i,0}}{d_{i,1}}$ is recorded. All items with a q_i above a defined percentile p are assigned to class 1, and the items below to class 0. By choosing a percentile threshold well above 50 the skewed distribution of class 0 and class 1 cases for both boundaries and accents can be tackled, i.e. more items receive class 0 than class 1. Other flat clustering approaches as kMeans assume equal variance of all classes and thus perform worse on the given skewed distributions. The tuned percentile thresholds p amount 87 and 82 for boundary placement and accentuation, respectively.

In a subsequent post-selection step for words with more than one accent assignment only the accent closest to the class 1 centroid is kept, and all other accents are removed.

6 Validation

Chunking, syllable extraction, phrase boundary assignment, and accent assignment were validated against the reference data in terms of F1, precision, recall, and accuracy. The best results for the tuning subset and the final results for the entire data set are shown in Table 1. Precision, recall, and F1-score are displayed for the class 1 cases only (i.e. *presence of boundary, accented*; values would be higher, if averaged over both classes).

Chunking was evaluated indirectly via measuring the precision of pause detection. In the lack of any manual reference the pauses were validated against the pause segments in the *MAU* segment tier. Since MAuS alignment also accounts for short within-chunk pauses not to be extracted for the given chunking task, only the precision will be reported. For this purpose, each extracted pause overlapping with exactly one *MAU* pause was counted as a true positive – exactly one, since two overlapping pauses indicate that a lexical item between these two pauses had been erroneously ignored.

Syllable nucleus extraction, again, in the lack of a manual reference, was evaluated against the MAuS segmentation. Precision, recall and F1 were calculated from the comparison of the syllabifier time stamps with the *MAU* syllable nuclei midpoints. Due to noise in the automatic segmentation a nucleus co-occurrence within a catch window of 0.1s was counted as a true positive.

Prosodic boundary assignment was evaluated on the word level against the manual annotations. Boundary decisions were to be made for each word boundary which was located based on the MAuS segmentation. To cope with misalignments between the segmentation and the manual prosodic event placements again a catch window of length 0.15s was defined.

Accent assignment was evaluated on the syllable level against the manual annotations using a catch window of length 0.1s.

For boundary and accent evaluation Table 1 additionally contains the baseline accuracies from assigning the most frequent class only, i.e. to label each word boundary with "no prosodic boundary" and each syllable with "not accented".

Table 1 – Results for chunking, syllable nucleus, boundary, and accent detection for the tuning data subset and for the entire data set. For boundary localization *Acc* and *BL* (both in %) refer to word accuracies. For accent localization they refer to syllable accuracies. The model underlying *BL* is given by the uniform assignment of the most frequent class (i.e. "no boundary" and "not accented", respectively).

	tuning data				all data				
	F1	**Prec**	**Rec**	**Acc**	**F1**	**Prec**	**Rec**	**Acc**	**BL**
Chunking	–	0.98	–	–	–	0.97	–	–	–
Syllables	0.88	0.94	0.83	–	0.87	0.93	0.82	–	–
Boundaries	0.61	0.64	0.58	84.3	0.59	0.67	0.53	85.0	79.2
Accents	0.63	0.56	0.70	71.1	0.59	0.54	0.66	70.5	67.4

7 Discussion

Quite good performances were obtained for chunking and syllable nucleus extraction. Furthermore, the other components' performances beat the corresponding baselines, which are already quite high due to the skewed boundary and accent distributions. This skewness makes it necessary to report not only accuracies but also F1-scores, precision, and recall. The obtained F1-scores are rather at the bottom end of the scores reported in [9, 10, 8] ranging from 0.58 up to 0.86. This can to a large extent be explained by differences in the underlying data. In contrast

to the mentioned studies that examined read news speech, the current approach was applied to spontaneous dialog speech which contains more variation due to irregular boundaries and hesitations. Further difficulties arise from the observation that several prosodic structure cues are less salient in spontaneous speech than in read speech, e.g. pitch reset as shown in [22]. Thus a comparative evaluation is not possible at the current stage.

In the exploratory phase four parameters were tuned for syllable nucleus extraction, and one parameter (the clustering percentile threshold) each for boundary and accent detection. For syllable detection the optimized values are sufficiently close to the values previously obtained for another data set of hand-segmented read speech by Nelder-Mead optimization [16]. This indicates good generalization capabilities across corpora. The percentile thresholds in contrast are expected to be more dependent on speaking-style related densities of phrase boundaries and accents and thus might need to be re-adjusted for other data sets.

A shortcoming of the current approach is its vulnerability to inherited errors due to the dependencies of processing steps on the outcome of the preceding steps. To give an example, accents can only be placed, where syllable nuclei have been detected, and several accent-related features are measured relative to the prosodic phrases extracted before. It is thus to be tested whether a disentanglement of the feature sets would increase performance. In addition, alternative cluster centroid bootstrapping assumptions e.g. referring to word predictability will be examined.

For all feature extraction and for the prosodic annotation the CoPaSul toolkit [23] was used. It is written in Python3 and is freely available here: [24].

8 Acknowledgments

The work of the author is financed by a grant of the Alexander von Humboldt society.

References

[1] GRICE, M., M. REYELT, R. BENZMÜLLER, J. MAYER, and A. BATLINER: *Consistency in Transcription and Labelling of German Intonation with GToBI*. In *Proc. ICSLP*, pp. 1716–1719. New Castle, Delaware, 1996.

[2] WIGHTMAN, C.: *ToBI Or Not ToBI?* In *Proc. Speech Prosody*, pp. 25–29. Aix-en-Provence, 2002.

[3] WIGHTMAN, C. and M. OSTENDORF: *Automatic labeling of prosodic patterns*. *IEEE Transactions on Speech and Audio Processing*, 2(4), pp. 469–481, 1994.

[4] SYRDAL, A., J. HIRSCHBERG, J. MCGORY, and M. BECKMAN: *Automatic ToBI prediction and alignment to speed manual labeling of prosody*. *Speech Communication*, 33(1–2), pp. 135–151, 2001.

[5] CHEN, K., M. HASEGAWA-JOHNSON, and A. COHEN: *An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic prosodic model*. In *Proc. of ICASSP*, pp. 1509–1512. 2004.

[6] SCHWEITZER, A. and B. MÖBIUS: *Experiments in Automatic Prosodic Labeling*. In *Proc. Eurospeech*, pp. 2515–2518. Brighton, 2009.

[7] ROSENBERG, A.: *AuToBI – a tool for automatic ToBI annotation*. In *Proc. Interspeech*. 2010.

[8] SUNI, A., J. ŠIMKO, D. AALTO, and M. VAINIO: *Hierarchical representation and estimation of prosody using continuous wavelet transform*. *Computer, Speech, and Language*, pp. 1–14, 2016.

[9] ANANTHAKRISHNAN, S. and S. NARAYANAN: *Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling*. In *Proc. of ICSLP*, pp. 297–300. 2006.

[10] JEON, J. and Y. LIU: *Automatic prosodic event detection using a novel labeling and selection method in co-training*. *Speech Communication*, 54, pp. 445–458, 2012.

[11] SCHIEL, F.: *Speech and speech-related resources at BAS*. In *Proc. 1st LREC*, pp. 343–349. Granada, Spain, 1998.

[12] REICHEL, U., F. SCHIEL, T. KISLER, C. DRAXLER, and N. PÖRNER: *The BAS Speech Data Repository*. In *Proc. LREC*, pp. 786–791. Portorož, Slovenia, 2016.

[13] SCHIEL, F.: *Automatic Phonetic Transcription of Non-Prompted Speech*. In *Proc. ICPhS*, pp. 607–610. San Francisco, 1999.

[14] BOERSMA, P. and D. WEENINK: *PRAAT, a system for doing phonetics by computer*. Tech. Rep., Institute of Phonetic Sciences of the University of Amsterdam, 1999. 132–182.

[15] SAVITZKY, A. and M. GOLAY: *Smoothing and Differentiation of Data by Simplified Least Squares Procedures*. *Analytical Chemistry*, 36(8), pp. 1627–1639, 1964.

[16] REICHEL, U.: *Linking bottom-up intonation stylization to discourse structure*. *Computer, Speech, and Language*, 28, pp. 1340–1365, 2014.

[17] NELDER, J. and R. MEAD: *A simplex method for function minimization*. *Computer Journal*, 7, pp. 308–313, 1965.

[18] PFITZINGER, H., S. BURGER, and S. HEID: *Syllable Detection in Read and Spontaneous Speech*. In *Proc. ICSLP*, vol. 2, pp. 1261–1264. Philadelphia, 1996.

[19] REICHEL, U. and K. MÁDY: *Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian*. In *Proc. Interspeech 2014*, pp. 111–115. Singapore, 2014.

[20] BEŇUŠ, V., U. REICHEL, and K. MÁDY: *Modelling accentual phrase intonation in Slovak and Hungarian*. In *Complex Visibles Out There*, vol. 4, pp. 677–689. Palacký University, Olomouc, Czech Republic, 2014.

[21] LEVOW, G.: *Unsupervised and semi-supervised learning of tone and pitch accent*. In *Proc. HLT-NAACL*, pp. 224–231. 2006.

[22] SWERT, M., E. STRANGERT, and M. HELDNER: *F0 declination in read-aloud and spontaneous speech*. In *Proc. ICSLP*, pp. 1501–1504. Philadelphia, PA, 1996.

[23] REICHEL, U.: *CoPaSul Manual – Contour-based parametric and superpositional intonation stylization*. RIL, MTA, Budapest, Hungary, 2016. Https://arxiv.org/abs/1612.04765.

[24] REICHEL, U. D.: *CoPaSul toolkit*. http://clara.nytud.hu/~reichelu/#p_sof, 2016. Version 0.2, December 30th, 2016.