Automatic labeling of prosody

Automatyczna anotacja prozodii

Agnieszka Wagner

Institute of Linguistics, Adam Mickiewicz University
wagner@amu.edu.pl

ABSTRACT

This paper proposes a framework for the automatic labeling of prosody which involves the detection and classification of pitch accents and phrase boundaries. Four statistical models are designed to perform these tasks on the basis of a compact and simple phonetic representation of prosody. The models yield high average accuracy and consistency and their output can be used to provide speech corpora with prosodic information.

STRESZCZENIE

Niniejsza praca przedstawia metodę automatycznej anotacji prozodii, na którą składa się detekcja i klasyfikacja akcentów i granic fraz intonacyjnych. Do realizacji tych zadań, na podstawie zwięzłej i prostej fonetycznej reprezentacji prozodii, zaprojektowano cztery modele statystyczne. Modele uzyskują wysoką średnią dokładność i zgodność, a dostarczone przez nie dane mogą zostać wykorzystane do uzupełnienia korpusów mowy o informację prozodyczną.

1. Introduction

Over the last two decades numerous approaches to the automatic labeling of prosody have been proposed (e.g. [1–6]), which is related to the growing interest in speech technologies such as speech synthesis or recognition. All types of speech applications rely on speech corpora which have to be provided with appropriate annotation at the segmental and suprasegmental level, the latter including more or less precise information concerning utterance prosody. The most important communicative functions of prosody include *prominence* and *phrasing*.

1.1. Functional aspects of prominence

Prominence can be attributed to the occurrence of a pitch accent associated with a metrically strong, stressed syllable. Pitch accents have a prominence-cueing function i.e., they cause a word or its part to stand out from its environment [7]. Prominence is closely related to information structure; various pitch accent types associated with prominent syllables convey considerable differences in meaning e.g., a fall accent (HL) introduces an entity into the background or shared knowledge of the interlocutors, whereas a fall-rise (HLH) is used to select an entity from the background. Thus, the use of a specific accent pattern determined by accenting some words and failing to accent

others, together with realization of specific pitch accent types, serves communication purposes.

Generally, there is no agreement as regards acoustic correlates of stress and prominence. In some studies [7, 8] variation in fundamental frequency (or pitch movements) and (to lesser extent) overall intensity are identified as the main cues signaling accentual prominence, while duration and spectral emphasis play a secondary role in this respect [9].

1.2. Functional aspects of phrasing

Phrasing organizes an utterance into a hierarchical prosodic structure. Intonation phrases include one obligatory (nuclear) pitch accent and are characterized by semantic and syntactic coherence. They constitute the domain of recurring intonation patterns and can be considered as units of information. Apart from a binary distinction between boundary presence vs. absence intonation phrase boundaries are classified with respect to *strength* (minor vs. major phrase boundary) and type (rising vs. falling). This kind of phrasing information can be used to resolve ambiguous parses, to disambiguate the meaning that can be assigned to a given phrase by the hearer or to enhance the performance of the language model component of ASR or dialogue systems..

Phrase boundaries are signaled mainly by timing cues – duration of syllables and vowels increases significantly in the vicinity of phrase boundaries [10, 11]. The significance of the silent pause in signaling phrase boundaries was shown in [12]. Pause duration is an important cue to boundary strength, but in the prediction of upcoming boundaries listeners use it only in the absence of a distinct pre-boundary lengthening (cf. [11]). Some studies also point out the role of F0 cues (e.g., [13]) and voice quality (e.g., [14]) in signaling phrase boundaries.

2. Features of the current approach

The objective of this study is to provide means for an efficient and effective automatic labeling of prosody. For this purpose we design models performing detection and classification of pitch accents and phrase boundaries. The minimum requirement is that the models achieve accuracy similar to that reported in other studies and comparable to levels of agreement among human labelers in manual transcription of prosody. Features of our framework of automatic intonation labeling can be summarized follows:

- We use acoustic and lexical features (cf. [1], [5]).
- Our acoustic feature vectors are simple in terms of extraction, because they can be easily derived from utterance's F0, timing cues and lexical features and exclude intensity (cf. [2, 3, 15]). With few exceptions we use features which refer to relative values.
- The acoustic features can be regarded as correlates of accentual prominence and phrase boundaries, and describe realization of different pitch accent and boundary tone types. The features constitute a phonetic representation which reflects melodic properties of intonation. From this representation a higher-level acoustic-perceptual description can be easily and reliably derived.

- Pitch accents and phrase boundaries are detected at the word level (cf. [1, 2, 5, 16]).
- In the automatic detection and classification of phrase boundaries both major and minor (intermediate) phrase boundaries are considered (cf. [1, 2, 4, 12, 16]).
- Our framework of prosody labeling consists of four statistical models, each using a different feature set and performing one of the tasks: detection of accentual prominence and phrase boundary position, classification of pitch accents and boundary tones (performed on syllables marked as being associated with a pitch accent or phrase boundary).
- For building the detection and classification models we apply neural networks, decision trees and discriminant function analysis.

3. Speech material and feature extraction

3.1. Speech corpus

The speech material comes from the corpus designed for the Polish module of the BOSS unit selection TTS system. The corpus contains recordings of phonetically rich and balanced sentences, fragments of fiction and reportage read in a reporting style by a professional male speaker. The whole speech material was automatically phonetically transcribed and segmented at the phoneme, syllable and word level. Stress was marked with the help of a large pronunciation lexicon. The position and types of pitch accents and phrase boundaries were annotated manually. The subset of the corpus used in the current study consists of 1052 utterances (15566 syllables).

3.2. Data preparation for analyses

3.2.1. F0 extraction and processing

F0 extraction and processing was performed with a Praat script. F0 was extracted every 10 ms (based on the autocorrelation method); all F0 values detected below and above thresholds describing speaker's range were treated as missing. In order to eliminate pitch perturbations and segmental effects the F0 contours were smoothed with a low-pass filter (5 Hz bandwidth). The unvoiced regions were interpolated. The waveforms were resynthesized with the smoothed and interpolated F0 contours using PSOLA.

3.2.2. F0 contour parametrization

In order to analyze the phonetic realization of pitch accents and phrase boundaries for each syllable and its vocalic nucleus features describing variation in F0 and duration were automatically extracted with a Praat script. For each syllable and vowel the following features were determined:

- F0 value at the start and end of the syllable and vowel,
- maximum, minimum and mean F0,
- timing of the F0 maximum and minimum,
- amplitude, slope, steepness and duration of the rising and falling pitch movement,
- *Tilt* parameter describing the shape of the pitch movement,
- start, end and overall duration.

On the basis of this representation new features were derived. Some of them referred to relative values of the parameters listed above, whereas others described pitch variation in a two-syllable window including the current and the next syllable or, alternatively, the previous syllable. For each syllable and vowel in the database features of the two previous and two next syllables/vowels were provided as well.

4. Determination of accentuation and phrasing

4.1. Acoustic correlates

In a series of extensive ANOVA and discriminant function analyses the effect of pitch accent and phrase boundary presence vs. absence on variation in the F0 and duration parameters extracted from the speech data was investigated. The goal was to find parameters which can be considered as acoustic correlates of pitch accents and phrase boundaries. Apart from taking into account statistically significant effects, the final feature set was determined in such a manner as to provide a phonetic representation of pitch accent and phrase boundary realization that is compact, has low redundancy and wide coverage. The representation is used to train models to detect pitch accents and phrase boundaries.

4.1.1. Pitch accents

Five features describing variation in F0 and duration were identified as acoustic correlates of pitch accents. They include (numbers in brackets show values of the F statistics):

- *slope*: a measure of an overall pitch variation on the syllable (F=902.65),
- relative syllable (F=770.27) and nucleus *duration* (F=489.45) calculated as in [3],
- *Tilt* parameter (F=648.47) describing the shape of the pitch movement on the syllable: –1 indicates falling pitch movement, 1 indicates rising movement and 0 indicates the same amount of rise and fall,
- height of *F0 peak* on the syllable (F=591.4).

The results of our analyses showed that accented syllables are characterized by almost twice as high average slope as unaccented syllables (131.18 vs. 73.39 Hz/s), which indicates that accentual prominence involves greater pitch variation. Prominent syllables also have higher F0 peaks (125.74 vs. 115.4 Hz) and higher value of the Tilt parameter (0.11 vs. –0.41). This effect shows that most pitch accents are realized by both rising and falling pitch movement and that pitch rises are generally better cues to prominence then pitch falls. Accented syllables/vowels are much longer than the stressed unaccented ones – the average difference is 20ms/10ms for syllables and vowels respectively (cf. Figure 1).

The effects discussed here are statistically significant ($p<0.01$). With one exception (i.e. relative syllable and nucleus duration) the features are not significantly correlated with one another, which ensures low redundancy of the resulting representation of accentual prominence.

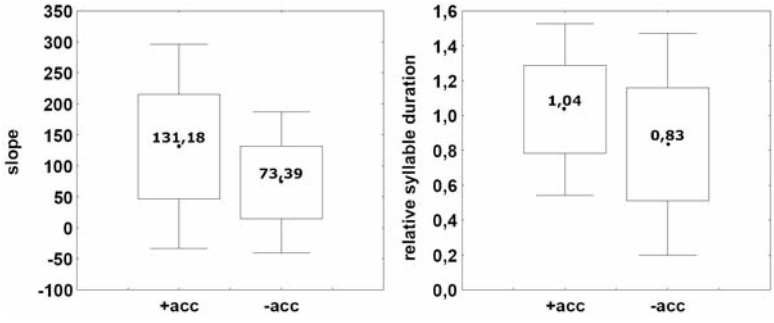

Figure 1: Mean slope (left) and relative duration (right) of accented (+acc) and stressed but unaccented syllables (-acc).

4.1.2. Phrase boundaries

In the time domain phrase boundaries are signaled most of all by:

- an increased duration of the pre-boundary syllable (+b=1.34) comparing to non-phrase-final syllable (-b=0.9, mean values, F=504.01),
- increased duration of the vowel of the word-penultimate syllable (+b=1.23 vs. –b=0.9, F=399.26),
- duration of the vowel of the pre-boundary syllable (+b=1.47 vs. –b=0.9, F=23.64),

In the F0 domain the most important features include:

- *Tilt* parameter (F=92.61) describing the shape of the pitch contour on the vowel,
- *F0mean* (F=81.07): overall F0 level on the vowel,
- *slope* (F=64.03) expressing the amount of pitch variation on the vowel of the word-penultimate syllable,
- *c1* (F=25.31): rising amplitude on the vowel

It was observed that vowels in syllables of a phrase-final position (+b) are characterized by significantly less falling pitch (indicated by higher average tilt value)

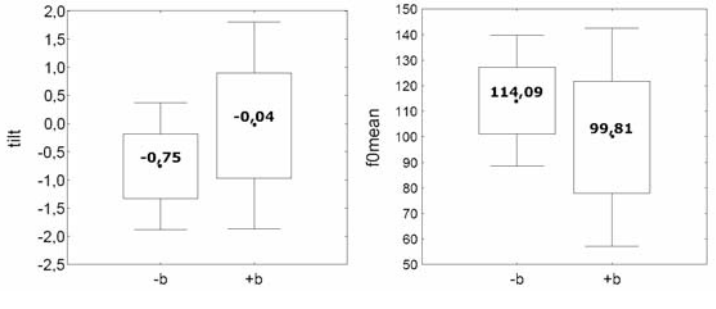

Figure 2: Mean Tilt and F0mean of vowels in pre-boundary (+b) and word-final syllables (–b).

than vowels in word-final syllables which do not precede a boundary (-b). This effect (cf. Figure 2) can be attributed to rising boundaries which signal continuation or interrogative mode.

Vocalic nuclei of phrase-final syllables have significantly higher rising amplitude (c1 +b=6.13) compared to other vowels (–b=0.38, mean values). The former are also characterized by significantly lower F0 mean, which can be attributed to falling boundaries (Fig. 2). The role of F0 features in signaling phrase boundaries is also confirmed by significantly greater amount of pitch variation (slope) on vowels in +b than –b class (129.4 Hz/s vs. 88.12 Hz/s). All the effects are statistically significant ($p<0.01$).

4.2. Automatic detection of accentual prominence

The detection of accentual prominence is performed at the word level i.e., only stressed syllables and vowels are considered. The models (decision tree, neural networks and DFA) were trained and tested on the subset of the Polish unit selection corpus consisting of 6417 stressed syllables divided into training (4278) and test (2139) samples. The ratio of accented to unaccented syllables was about 2:1 in each sample.

The best discrimination between accented (+acc) and unaccented stressed syllables (-acc) was achieved with neural networks: in the test sample the RBF (radial basis function) network performed the task with an average accuracy of 81.95 %, whereas the MLP network yielded an average accuracy of 81.72%. Slightly worse detection accuracy was observed for the classification tree and discriminant function analysis: 79.13% (test sample) and 77.23% (cross-validation test) respectively (Table 1).

Table 1: Results of accentual prominence detection at the word level (chance level accuracies are given in brackets, column class).

class	MLP	RBF	d. tree	DFA
+acc (61.18%)	81.79	81.76	77.06	74.89
–acc (38.82%)	81.65	82.14	81.2	80.94
Average (%):	81.72	81.95	79.13	77.92

4.3. Automatic determination of phrasing

The models performing the automatic detection of phrase boundary position were trained and tested on a subset of 6844 syllables of a word-final position including 1880 syllables followed by a phrase boundary. The highest average accuracy was achieved with discriminant function analysis (DFA) i.e., 82.05% (in the cross-validation test), but at the same time only 74.04% of phrase-final syllables were correctly identified using this method. The RBF network had the best performance in this respect, as it enabled correct identification of 81.55% of phrase boundaries (Table 2).

Table 2: Results of phrase boundary detection at the word level.

class	MLP	RBF	d. tree	DFA
–b (72.59%)	81.99	79.29	84.33	90.05
+b (27.14%)	76.26	81.55	78.6	74.04
Average (%)	79.13	80.42	81.47	82.05

5. Classification of pitch accents and phrase boundaries

5.1. The acoustic-perceptual representation of prosody

The acoustic-perceptual representation encodes melodic and functional aspects of prosody, and consists of an inventory of five pitch accent and five boundary tone types. Pitch accents are distinguished on the basis of melodic properties such as: direction, range and slope of the distinctive pitch movement, and its temporal alignment with the accented vowel. Pitch accents are described in terms of discrete bi-tonal categories: LH*, L*H, H*L, HL*, LH*L, where L indicates a lower and H a higher tonal target, and the asterisk indicates alignment with the accented vowel. Boundary tone types are distinguished on the basis of: direction of the distinctive pitch movement (rising vs. falling), amplitude of the movement, scaling of the f0 targets at the start/end of the movement and strength of the phrase break (intermediate/minor vs. major phrase boundary, e.g. 2,? stands for a minor rising boundary and 5,. for a major falling boundary).

5.2. Classification of pitch accents

The models designed for pitch accent classification rely on a small vector consisting of eight features selected from among the parameters derived from acoustical and lexical features of the utterance on the basis of statistically significant effects of different pitch accent types. The features include: *amplitude* of the rising/falling F0 movement, *relative mean*, *maximum* and *minimum F0* determined for the vocalic nucleus and Tilt, Tilt amplitude and direction (calculated as a ratio of mean F0) determined in a two-syllable window containing accented syllable and the next syllable. The resulting representation is compact, has low redundancy and constitutes part of the phonetic description of intonation. The most significant effects of pitch accent type on variation in two acoustic parameters *Tilt* and *amplitude of the falling F0 movement* are depicted in Figure 3.

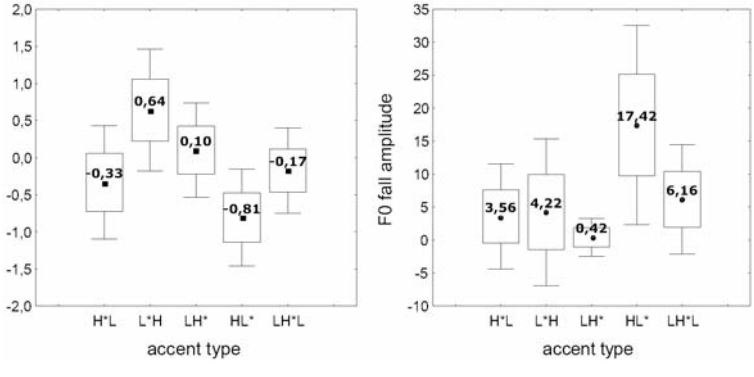

Figure 3: Variation in the average values of Tilt and falling amplitude parameters depending on pitch accent type.

The models were trained on a subset of 3671 syllables marked as being associated with a pitch accent. The distribution of pitch accents among the five categories was unequal: there were 1401 instances of H*L accents and only 96 instances of LH*L accents. The syllables were proportionally divided into a training (2754) and test sample (917).

Table 3. Results of pitch accent type classification

class	d.tree	DFA	MLP	RBF
H*L (36.86%)	71.01	71.89	76.63	78.99
L*H (10.25%)	86.17	70.21	77.66	70.21
LH* (28.9%)	70.19	80.38	83.02	85.66
HL* (20.94%)	91.67	88.54	89.58	89.58
LH*L (3.05%)	89.29	85.71	60.71	39.29
Average (%):	81.67	79.36	77.52	72.75

The performance of the classification models summarized in Table 3 shows that the results are much better than a chance level pitch accent type recognition (values in column *class*) in which most of the pitch accented syllables are labeled as H*L. It means that the features used by the models discriminate well among different pitch accent types and proves that the resulting representation has wide coverage. The decision tree outperformed the other models and yielded an average accuracy of 81.67% (Table 3).

5.2.1. Classification of boundary tones

The models designed for boundary tone classification rely on three F0 features and one lexical feature:

- *F0end*: F0 level at the end of the word-final syllable,
- *F0mean*: overall F0 level on the nucleus of the word-penultimate syllable,
- *direction*: direction of the distinctive pitch movement at the edge of the phrase,
- *distance to the next pause*: distinction between boundaries of a different strength.

These features were selected on the basis of statistically significant variation due to different boundary tone types. They constitute part of the phonetic description of prosody which is characterized by wide coverage and low redundancy.The most significant effects of phrase boundary type on the acoustic parameters (*F0end* and *F0mean*) are depicted in Figure 4.

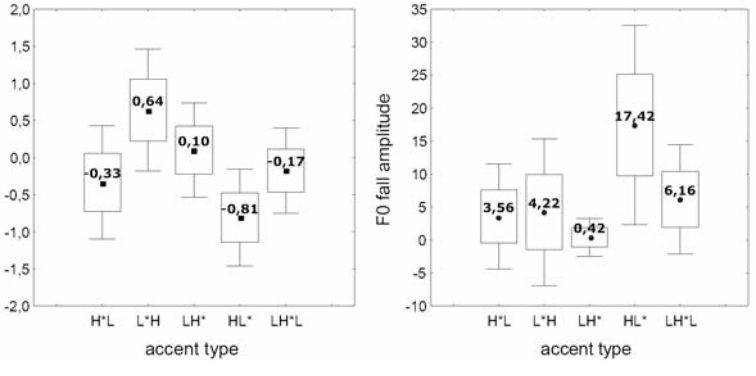

Figure 4: Variation in the average values of F0mean (left) and F0end (right) depending on phrase boundary type: 2,. – minor falling, 2,? – minor rising, 5,. – major falling and 5,? – major rising.

The models for boundary tone classification were trained on a subset of 1502 syllables marked as being associated with a phrase boundary. One category of falling boundary tones was excluded from the classification due to data sparseness. The performance of the models is summarized in Table 4.

Table 4. Results of boundary tone classification

class	d. tree	DFA	MLP	RBF
2.. (20.42%)	70.13	70.13	66.23	70.13
2.? (33.42%)	79.37	70.63	80.16	84.92
5.. (37.93%)	92.25	99.3	98.6	98.6
5.? (8.22%)	96.77	83.87	93.55	96.77
Average (%):	84.63	80.98	84.64	87.61

It can be seen that the models perform significantly better than a chance-level boundary type recognizer (values in column *class*). The average accuracy yielded by the models varies between 80.98% (DFA) and 87.61% (RBF network). The average recognition accuracy achieved with the classification tree and MLP network is above 84%. Generally, weak boundaries (labeled 2) were more difficult to recognize than strong boundaries (labeled 5). The average recognition accuracy of the former did not exceed 85%, whereas the latter were recognized with at least 92% accuracy.

6. Discussion and final remarks

The objective of the current study was to propose a framework of an efficient and effective automatic labeling of prosody. The labeling involves detection and classification of pitch accents and phrase boundaries. Each task is performed by a different model (decision tree, DFA, MLP or RBF network) and the models rely on different feature sets which constitute the phonetic representation of prosody. This representation is compact as it consists of 23 features altogether and simple – it can be easily derived from the utterance's acoustics and lexical features. It also has wide coverage, because it enables an accurate and consistent detection of accentual prominence and phrase boundaries, as well as an efficient recognition of pitch accent and boundary tone types distinguished at the acoustic-perceptual level. Low redundancy of the phonetic representation is ensured by the fact that its components can not be derived from one another.

Generally, the models proposed in our framework perform significantly better than a chance-level detector which assigns the most frequent label to all syllables. As regards accentual prominence detection the models achieve accuracy similar to that reported in other studies ([2], [12], [15]) and approaching the levels of agreement between human labelers ([17], [18]). The best model designed for pitch accent type classification (decision tree) yielded an average accuracy of 81.67%, which compares favorably with ([12], [15]). The models for phrase boundary detection outperformed [1] where the average accuracy was about 71% and yielded accuracy similar to [11]. As regards automatic boundary type classification the results are similar to that achieved by the best models ([4], [6], [12]).

The advantage of the framework proposed in the current study is the use of a compact and simple representation of utterance prosody as a basis for detection and classification tasks performed by models which yield accuracy comparable to that reported in other studies and approaching the levels of agreement among human annotators in the manual labeling of prosody.

Acknowledgement

This work is supported from the financial resources for science in the years 2010–2012 as a development project (National Centre for Research and Development).

References

[1] Wightman, C.W. and Ostendorf, M. 1994. Automatic Labeling of Prosodic Patterns. In *IEEE Trans. Speech and Audio Proc.*, 4(2): 469-–481.

[2] Kießling, A., Kompe, R., Batliner, A., Niemann, H. and Nöth, E. 1996. Classification of Boundaries and Accents in Spontaneous Speech. In *Proc. 3rd CRIM/FORWISS Workshop, Montreal 1996*, pp. 104–113.

[3] Rapp, S. 1998. Automatic labeling of German prosody. In *Proc. ICSLP, Sydney 1998*, pp. 1267–1270.

[4] Wightman, C.W., Syrdal, A., Stemmer, G., Conkie, A. and Beutnagel, M. 2000. Perceptually Based Automatic Intonation labeling and Prosodically Enriched Unit Selection Improve Concatenative Text-To-Speech Synthesis. In *Proc. ICSLP, Beijing 2000*, pp. 71–74.

[5] Ananthakrishnan, S. and Narayanan, S. S. 2008. Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. In *IEEE Trans. Speech and Audio Proc.*, 16(1): 216–228.

[6] Schweitzer, A. and Möbius, B. 2009. Experiments on automatic prosodic labeling. In *Proc. Interspeech 2009, Brighton*.

[7] Terken, J. 1991. Fundamental frequency and perceived prominence of accented syllables. In *J. Acoust. Soc. Am.*, (89): 1768–1776.

[8] Jassem, W. 1961. *Accent of Polish*, Polish Academy of Sciences, Kraków.

[9] Sluijter, A. M. C. and van Heuven, V. J. 1996. Acoustic correlates of linguistic stress and accent in Dutch and American English. In *Proc. ICSLP 1996*, pp. 630–633.

[10] Yoon, T.J., Cole, J. and Hasegawa-Johnson, M. 2007. On the edge: Acoustic cues to layered prosodic domains. In *Proc. 16th Int. Cong. Phon. Sci., Saarbruecken 2007*, pp. 1017–1020.

[11] Aguilar, L., Bonafonte, A., Campillo, F. and Escudero, D. 2009. Determining Intonational Boundaries from the Acoustic Signal. In *Proc. INTERSPEECH 2009, Brighton 2009*.

[12] Bulyko I. And Ostendorf M. 2001. Joint prosody prediction and unit selection for concatenative speech synthesis. In *Proc. ICASSP, Salt Lake City 2001*, pp. 781–784.

[13] Carlson, R., Hirschberg, J. and Swerts, M. 2005. Cues to upcoming Swedish prosodic boundaries: subjective judgment studies and acoustic correlates. In *Speech Comm*. 46 (3/4): 326–333.

[14] Carlson, R. and Swerts, M. 2003. Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials. *Proc. 15th Int. Congr. Phonet. Sci., Barcelona 2003*, pp. 507–510.

[15] Sridhar, R., Nenkova, A., Narayanan, S.S. and Jurafsky, D. 2008. Detecting prominence in conversational speech: pitch accent, givenness and focus. In *Proc. Speech Prosody, Campinas 2008*, pp. 453–457.

[16] Ross, K. and Ostendorf, M., "Prediction of abstract prosodic labels for speech synthesis. In *Computer Speech and Language*, (10): 155–185, 1996.

[17] Pitrelli, J.F., Beckman, M.E. and Hirschberg, J. 1994. Evaluation of prosody transcription labeling reliability in the ToBI framework. In *Proc. ICSLP, Yokohama 1994*, pp. 123–126.

[18] Grice, M., Reyelt, M., Benzmueller, R., Mayer, J. and Batliner, A. 1996. Consistency in transcription and labeling of German intonation with GToBI. In *Proc. ICSLP, Philadelphia 1996*, pp. 1716–1719.