Automatic labeling of intonation using acoustic and lexical features

Conference Paper · June 2009

1 author:

Agnieszka Wagner
Adam Mickiewicz University
36 PUBLICATIONS **175** CITATIONS

SEE PROFILE

Automatic Labeling of Intonation Using Acoustic and Lexical Features

Agnieszka Wagner

Department of Phonetics, Institute of Linguistics
Adam Mickiewicz University in Poznań, Poland
wagner@amu.edu.pl

Abstract

This paper proposes a framework of automatic intonation labeling which involves 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 representation consisting of features identified as the main acoustic correlates of accentual prominence and phrase boundaries or describing the acoustic-phonetic realization of different types of pitch accents and boundary tones. The features can be easily derived from utterance's acoustics (F0 and timing cues) and lexical features. The models yield high average accuracy - between 80% and 87% depending on the task, and high consistency - they approach the levels of agreement among human labelers in manual transcription of intonation.

1. Introduction

Over the last two decades numerous approaches to automatic labeling of intonation have been proposed (e.g., [1], [2], [3], [4]), 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's intonation. The advantage of automatic intonation labeling is that it is significantly less laborious and time-consuming than manual transcription of intonation, but at the same time yields comparable accuracy and consistency (cf. [5], [6], [7]).

Most of the existing models performing the automatic intonation labeling rely on large vectors of features (up to 276 features in [8]) of a different type (e.g., acoustic, lexical, syntactic) which are considered as the main cues signaling pitch accents (accentual prominence) and phrase boundaries. Automatic classification of pitch accents and boundary tones requires identification of features which discriminate well between different categories (e.g., [2], [3]).

The statistical modeling techniques applied to automatic labeling of intonation include neural networks ([8], [9], [10]), classification trees ([2], [3], [11]), maximum entropy models [4], HMMs [12] or discriminant function analysis [9]. The models' complexity depends on speaking style (neutral, reporting style vs. conversational speech), the number of categories that need to be recognized (binary vs. multiple classification) and the number of tasks performed by the model e.g., instead of one model, two models can be designed, each to perform a different task like pitch accent or phrase boundary detection.

As regards performance of the models designed for automatic detection of accentual prominence they achieve an overall accuracy between 70% [3] and 87% [2]. In manual annotation of pitch accent position the inter-transcriber agreement is 80%-90% (e.g., [5], [6], [7]).

Better results are reported for automatic detection of phrase boundary position: an overall accuracy varies between 84% [4] and 94% [11]. Inter-transcriber agreement in this task exceeds 90% and is generally higher than for prominence marking (cf. [13]).

In the automatic classification of pitch accent types the best models yield accuracy of about 80% ([2], [9]). Pitch accent type identification by human labelers is performed with a lesser consistency than discrimination between accented and unaccented syllables, but there is a discrepancy in the agreement levels reported in different studies e.g. 64% in [5] and almost 95% in [7].

The accuracy of automatic boundary type recognition is much higher and sometimes exceeds 90% ([3], [11]). Joint recognition of pitch accent and boundary tone type yields accuracy of 78,6% [2]. Generally, the performance of phrase boundary type classifiers is comparable to levels of agreement among human labelers in this task i.e., 80%-90%.

2. Features of our approach

The objective of this study is to provide means of an efficient and effective automatic labeling of intonation. For that purpose we design models performing detection and classification of pitch accents and phrase boundaries. As regards performance 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 intonation. Features of our framework of automatic intonation labeling are as follows:

- We use acoustic and lexical features as opposed to ([4], [10]) where syntactic features are used as well or [1] that relies on acoustic features only.
- In terms of extraction our acoustic feature vectors are much simpler in comparison to those used in other studies, because they can be easily derived from utterance's F0, timing cues and lexical features, and exclude intensity features (cf. [2], [4], [8], [14]). With few exceptions we use features which refer to relative values i.e., values normalized against variation due to prosodic structure (in case of F0 features, cf. [15]) or caused by intrinsic properties of phonemes and especially vowels (in case of duration features).
- The set of lexical features differs from those used in other studies (e.g., [2], [4], [8]) and consists of: vowel type (distinction between short vs. long vowels), syllable

type (grouping based on nucleus type and composition of the onset and coda: voiced obstruent, obstruent or sonorant), lexical stress, distance to the next pause, syllable and word boundary. The lexical features are obtained fully automatically and the reliability of the extracted data is ensured by verification against a large pronunciation lexicon. With one exception (see sec. 5.3), the lexical features are not used directly by the models - they are necessary to derive the acoustic features.

- Our 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 of intonation which is compact (it is defined in terms of a small number of acoustic features), has low redundancy (its components can not be derived from one another), and wide coverage (it encodes distinctions between utterances which are perceptually different). Like other phonetic descriptions (e.g., [16], [17]) it reflects melodic properties of intonation. From this representation a higher-level acoustic-perceptual description of intonation can be easily and reliably derived. It is in terms of discrete tonal categories of pitch accents and boundary tones, and encodes melodic and functional aspects of intonation (see sec. 5.1).
- Pitch accents and phrase boundaries are detected at the word level as opposed to ([1], [10], [18]) where the detection is performed at the syllable level. We share the view expressed in [8] that for semantic or syntactic analysis it is more important to know which word rather than syllable is accented or phrase-final.
- In the task of automatic detection and classification of phrase boundaries both major and minor (intermediate) phrase boundaries are considered. In this regard our approach is similar to [10] and differs from majority of previous studies (e.g., [1], [3], [8], [11], [18]).
- Contrary to [3] the framework proposed in the current study consists of four statistical models, each using a different feature set and performing one of the tasks: detection of accentual prominence, detection of phrase boundary position, classification of 5 pitch accent types and recognition of 4 boundary tone types. The two latter tasks are 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 (NN), decision trees and discriminant function analysis (DFA). Neural networks (multilayer perceptrones - MLP and radial basis function networks - RBF) are trained using the back-propagation algorithm and/or conjugate gradient descend method. For each task models of various complexity and trained with various prior classification probabilities are tested, and the most efficient ones are selected for further optimization. As regards decision trees the QUEST classification tree program [19] is used. The models are designed semi-automatically using Statistica 6.0 [20].

It is assumed that the features of our automatic intonation labeling framework ensure its effectiveness and an efficient application in the field of speech technology.

3. Data corpus and features

3.1. The Polish unit selection corpus

The speech corpus used in the current study was built for the Polish module of BOSS (Bonn Open Source Synthesis) unit selection speech synthesis system [21]. The corpus contains 4 hours of recordings of a professional male speaker reading phonetically rich and balanced sentences, fragments of fiction and reportage, dialogues, and texts from railway information. From this corpus a subset consisting of 1052 utterances representative of the whole speech material was selected and annotated at the segmental and suprasegmental level.

3.1.1. Segmental annotation

Phonetic transcription was performed automatically using an inventory of 39 phonemes for broad transcription [22]. The computer coding conventions were based on SAMPA for Polish and IPA alphabet. The phonetic segmentation was performed by a program CreatSeg which uses HMM models and yields accuracy of about 80-90% depending on the type of transcription (broad vs. narrow).

3.1.2. Suprasegmental annotation

The information above the segmental level included: lexical stress, syllable boundaries, word boundaries (with the distinction between orthographic and prosodic words), pause markers and voice quality (vocal fry or creaky voice). Except for the latter one, all this information was obtained automatically. Boundaries and stress markers were verified with the help of a large pronunciation lexicon. The utterances were also manually labeled with intonational features: accent types, strength of a phrase break and tune type were marked on the basis of a representation proposed in [23].

3.2. F0 extraction and processing

As automatic extraction of pitch is prone to errors (mostly due to microprosody or voice quality) some preprocessing is necessary in order to get F0 data which can be regarded as a reliable basis for parameterization and analysis. For the purpose of the current study a Praat (version 4.1.5) script was written which performed the extraction and preprocessing of F0. Pitch was extracted every 10ms (based on autocorrelation method) and all values detected below and above the pitch range (65Hz-220Hz) were treated as missing. In order to eliminate pitch perturbations and segmental effects smoothing of the F0 contours with a low-pass filter (5Hz bandwidth) was performed. The unvoiced regions were interpolated through and waveforms were resynthesized with the smoothed and interpolated F0 contours using PSOLA. The visual inspection as well as perceptual analysis of the resulting F0 contours showed that the extraction and preprocessing made it possible to eliminate short-term deviations and at the same time to preserve the intonational features.

3.3. F0 contour parameterization

In order to analyze the acoustic-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; they included:

- F0 value at the start/end of the syllable/vowel,
- maximum/minimum and mean F0,
- timing of the F0 maximum/minimum,
- amplitude/slope/steepness/duration of the rising/falling pitch movement calculated as in [17],
- Tilt parameter describing the shape of the pitch movement calculated as in [16],
- start/end and overall duration of a syllable/vowel

From the resulting 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 the current and 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 this section the results of extensive analyses carried out in [15] and aiming at identification of acoustic correlates of pitch accents and phrase boundaries are presented. The analyses (ANOVA, discriminant function analysis - DFA) investigated the effect of a pitch accent and phrase boundary presence vs. absence on variation in the F0 and duration parameters extracted from the speech data. Apart from taking into account statistically significant effects, the final feature set was determined in such a manner as to provide a compact representation of pitch accent and phrase boundary realization at the acoustic-phonetic level characterized by low redundancy and wide coverage. The representation is used to train models to detect pitch accents and phrase boundaries.

4.1.1. Pitch accents

A shown in ([9], [24], [25]) accentual prominence is signaled primarily by variation in fundamental frequency, whereas variation in intensity, duration and spectral emphasis play a secondary role in this respect (e.g., [26]). For the purpose of the current study the five features listed below were identified as acoustic correlates of pitch accents:

- *slope* describing overall pitch variation on a syllable (F=902,65)
- *relative syllable* (F=770,27) and *nucleus duration* (F=489,45) calculated as in [2]
- *Tilt* (F=648,47) describing the shape of the pitch movement on a syllable
- *F0max* - height of F0 peak on a syllable (F=591,4)

All the features are significantly affected by a pitch accent presence/absence ($p<0,01$). ANOVA results (values of the F statistics are given in brackets) indicate that variation in F0 distinguishes accented from unaccented syllables to a greater extent than variation in duration. With one exception (relative syllable and nucleus duration) the features are not significantly correlated with one another, which ensures that the resulting representation has low redundancy.

4.1.2. Phrase boundaries

Pre-boundary lengthening belongs to the most salient cues signaling phrase boundaries (e.g., [27], [28]). Less agreement can be found as regards the reliability of phrase boundary detection using duration of the following silent pause (cf. [11], [28], [29]). Some studies (e.g., [30], [31]) point out the role of F0 cues and voice quality (vocal creak) in predicting upcoming phrase breaks.

In the current study a vector consisting of seven features which can be easily derived from F0, timing cues, and lexical features is used for detection of major and minor phrase boundary position. Except for *relative syllable duration* (F=504,01) the features describe variation in F0 or duration over the length of a vowel. They include: *relative duration of the current* (F=23,64) and *previous nucleus* (F=399,26), *Tilt* (F=92,61), *F0mean* (F=81,07), *slope* (F=64,03) and *amplitude* of the rising pitch movement (F=25,31).

All the features are significantly affected by presence of a phrase boundary ($p<0.01$) and can be regarded as its acoustic correlates. Together with the features identified as the main cues signaling accentual prominence they constitute part of the phonetic representation of intonation. The values of the F statistics (given in brackets) indicate that duration features are better cues to phrase boundary presence than F0 features. Except for relative syllable and nucleus duration, the features are not significantly correlated with one another, which ensures low redundancy of the resulting representation.

4.2. Detection of accentual prominence

The detection of accentual prominence is performed at the word level i.e., only stressed syllables/vowels are considered. The models (decision tree, NN 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) sample. The ratio of accented to unaccented syllables was about 2:1 in each sample.

The best discrimination between accented (marked as +acc) and unaccented stressed syllables (-acc) was achieved with neural networks: in the test sample the RBF network (82 radial neurons in the hidden layer) performed the task with an average accuracy of 81,95 %, whereas the MLP network (17 neurons in the hidden layer) 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 3 summarizes the average accuracy of correct detections of accented and unaccented syllables in the test sample (or cross-validation test in case od DFA) using different statistical modeling techniques. The numbers in brackets (column class) show chance-level accuracies.

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

Table 1: Results of accentual prominence detection at the word level (test sample).

It can be seen that the models designed in the current study

perform much better than a chance-level detector which assigns the most frequent label (here: +acc) to all syllables. They achieve accuracy similar to that reported in other studies ([8], [11], [14]) and to the levels of agreement between human labelers e.g., 80,6% in [5]. The best state-of-the-art models perform better and achieve accuracy between 84-87% ([2], [3], [4], [10]), but the advantage of our approach is the use of a compact and simple representation consisting of five features (as opposed to 276 features in [8]) which can be easily derived from utterance's acoustics and lexical features.

The results of sensitivity analysis conducted on the inputs to the neural networks (MLP and RBF) showed that features describing variation in F0 (tilt, F0max and slope) are more significant to the detection of accentual prominence than features reflecting variation in nucleus and syllable duration. This generally confirms our previous conclusions drawn on the basis of ANOVA results and the findings presented in the literature which indicate that variation in F0 is the main acoustic correlate of a pitch accent, whereas duration plays an important, but secondary role in this respect.

4.3. Detection of phrase boundaries

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. Among them there were 1880 syllables followed by a phrase boundary.

From among the three modeling techniques: NN, decision trees and DFA the latter achieved the highest average accuracy i.e., 82,05% (in the cross-validation test), but at the same time only 74,04% of phrase-final syllables (marked as +b) were correctly identified using this method. The best performance in this respect had the RBF network (with 23 neurons in the hidden layer) as it enabled correct identification of 81,55% of phrase boundaries. Apart from that the network yielded high average accuracy (80,42%) and performed well the identification of non-phrase-final (-b) syllables (79,29%). Table 4 summarizes the average accuracy of phrase boundary detection in the test sample (and cross-validation test in case of DFA) achieved by different models; chance-level accuracies are given in brackets.

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

Table 2: Results of phrase boundary detection at the word level (test sample).

It can be seen that the results are much better than a chance-level boundary detection in which the most frequent label (here: -b) is assigned to all syllables. The performance of the models proposed in the current study compares favorably with [1] where the average accuracy of phrase boundary detection was about 71%. But generally, our models perform worse than the best phrase boundary detectors which yield an overall accuracy well above 90% ([3], [10], [11]). The accuracy achieved by our models is also much below the levels of agreement among human labelers in manual identification of phrase boundaries which is 93%-98% for major intonational phrase boundaries. However, most of the existing approaches to prosodic boundary detection rely on large feature vectors including acoustic, lexical and syntactic features and more importantly - they take only major intonational phrase boundaries into account. On the contrary, our models enable detection of both major and minor phrase boundaries, which is the advantage of our framework. Moreover, they require only seven acoustic features to perform this complex task. For that reason the 80,42% average accuracy yielded by the RBF network can be regarded as satisfying and the use of a compact and simple representation of the acoustic-phonetic realization of phrase boundaries can be considered as another advantage of the framework proposed in the current study.

The results of sensitivity analysis carried out on the inputs to the RBF network confirm our previous conclusions drawn on the basis of ANOVA results and are in accordance with the findings presented in the literature. Namely, they show that in phrase boundary signaling features describing variation in F0 (tilt, slope, f0mean) play an important, but secondary role in comparison to features which reflect variation in duration.

5. Classification of accents and boundaries

In this section we present models performing an automatic classification of pitch accents and boundary tones into tonal categories which constitute a representation of intonation at the acoustic-perceptual level.

5.1. The acoustic-perceptual representation of intonation

The acoustic-perceptual representation of intonation used in the current study was proposed in [15]. It was derived from the intonation labeling framework applied in the Polish unit selection corpus (see sec. 3.1.2). The representation encodes melodic and functional aspects of intonation, and consists of an inventory of five pitch accent and five boundary tone types. The former are distinguished on the basis of melodic properties similar to those used in the IPO system [32]: 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 which of the two tones is aligned with the accented vowel. Boundary tone types are distinguished on the basis of: direction of the distinctive pitch movement (rise marked with ? vs. fall marked with a dot), amplitude of the movement, scaling of the f0 targets at the start/end of the movement and strength of the phrase break (2 marks intermediate phrase boundary, whereas 5 - major intonational phrase 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 utterance's acoustics and lexical features (see sec. 3.3) 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* [16] and *direction* (calculated as a ratio of mean F0) determined in a two-syllable window containing accented and the next syllable. The resulting representation is compact and has low redundancy (except for Tilt and Tilt amplitude no significant correlations among the

features were found) and constitutes part of the phonetic description of intonation.

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).

The table below summarizes the performance of different models i.e., it shows the percentage of correct classifications in the test sample and in the cross-validation test (in case of DFA). Chance-level accuracies are given in brackets.

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

Table 3: Results of pitch accent type recognition.

It can be seen that the results are much better than a chance-level pitch accent type recognition 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 (28 splits, 29 terminal nodes) performed better than the other models and yielded an average accuracy of 81,67%, which compares favorably with ([11], [18]). The advantage of our approach is high accuracy which exceeds the levels of agreement among human labelers in manual pitch accent type assignment reported in some studies e.g., [5] and the use of a simple and compact phonetic representation.

5.3. Classification of boundary tones

The models designed for boundary type classification rely on three F0 features - *syllable-final F0, mean F0* on the previous nucleus and *direction* which distinguish between boundary tones of a different direction and amplitude, and one lexical feature which describes *distance to the next pause* which distinguishes between boundaries of a different strength. They were selected on the basis of statistically significant variation due to different boundary tone types. The features constitute part of the phonetic description of intonation. There are no significant correlations among them, which ensures that the resulting representation has low redundancy.

The models (decision tree, DFA, MLP and RBF networks) were trained on a subset of 1502 syllables marked as being associated with a minor or major phrase boundary. One category of falling boundary tones was excluded from the classification due to data sparseness.

The performance of different models is summarized in the table below which shows the percentage of correct boundary tone type classifications in the test sample (or in the cross-validation test in case of DFA). Chance-level accuracies are given in brackets (column class).

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	99,30	98,6	98,6
5,? (8,22)	96,77	83,87	93,55	96,77
average%:	84,63	80,98	84,64	87,61

Table 4: Results of boundary tone type recognition.

It can be seen that the models perform significantly better than a chance-level boundary type recognizer. The average accuracy yielded by the models varies between 80,98% (DFA) and 87,61% (RBF network, 54 neurons in the hidden layer). The average recognition accuracy achieved with the classification tree (9 splits, 10 terminal nodes) and MLP network (20 neurons in the hidden layer) is above 84%. These results are comparable to the levels of agreement among human annotators in manual boundary tone type labeling which varies between 80% and 90%.

In general, weak boundaries (2,? and 2,.) were more difficult to recognize than strong boundaries (5,? and 5,.) The average recognition accuracy of the former did not exceed 85%, whereas the latter were recognized with at least 92% accuracy. These results are comparable to ([3], [11]) or better than those reported in other studies (e.g., [18]). High accuracy and consistency comparable to that obtained in manual boundary tone type labeling achieved using only four features (acoustic and lexical) can be regarded as an advantage of the framework proposed in the current study.

6. Final remarks

The objective of the current study was to propose a framework of an efficient and effective automatic labeling of intonation. 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 intonation. This representation is compact as it consists of 23 features altogether and simple, because it can be easily derived from 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.

The acoustic-perceptual description of intonation consists of an inventory of discrete tonal categories. It provides a more precise information on utterance's intonation, because it encodes both melodic and functional aspects of intonation. This information can be effectively used to disambiguate utterance's meaning [33] or to improve the output quality of the intonation generation component in the speech synthesis framework [34].

The advantage of our automatic intonation labeling framework is high average accuracy achieved in the detection and classification tasks while using a compact and simple representation. The models perform significantly better than a chance-level detector which assigns the most frequent label to

all syllables. In three out of four tasks the performance of the RBF network-based models is superior to that of other models. Only in pitch accent type recognition the decision tree outperforms other models. The accuracy yielded by the models is comparable to that reported in other studies and approaches the levels of agreement among human annotators in manual identification of position and types of pitch accents and phrase boundaries. All that ensures that our automatic intonation labeling is effective and efficient.

In the future the framework proposed in the current study will be applied to design models performing a speaker-independent automatic labeling of intonation.

7. References

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

[2] Rapp, S., "Automatic labeling of German prosody", in *Proc. ICSLP, Sydney 1998,* pp. 1267-1270

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

[4] Sridhar, R., Bangalore, V. K. and Narayanan, S. S., "Exploiting Acoustic and Syntactic Features for Automatic Intonation labeling in a Maximum Entropy Framework", *IEEE Trans. Speech and Audio Proc.,* 16(4):797-811, 2007.

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

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

[7] Yoon, T.J., Heejin, K., Chavarria, S. and Hasegawa-Johnson, M., "Inter-transcriber Reliability of Prosodic Labeling on Telephone Conversation Using ToBI", in *Proc. of ICSLP, Jeju 2004,* pp. 2729-2732

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

[9] Demenko, G., *Analysis of Polish suprasegmentals for needs of Speech Technology*, Adam Mickiewicz University Press, Poznań, 1999.

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

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

[12] Boidin C. and Boeffard O., "Modeling intonation variability with HMM for speech synthesis", in *Proc. Speech Prosody, Campinas 2008,* pp.115-119

[13] Mo, Y., Cole, J. and Lee, E.-K., "Naïve listeners' prominence and boundary perception", in *Proc. Speech Prosody, Campinas 2008,* pp. 735-739

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

[15] Wagner, A. "A comprehensive model of intonation for application in speech synthesis", PhD dissertation, Adam Mickiewicz University in Poznań, Poznań, 2008.

[16] Taylor, P., "Analysis and synthesis of intonation using the tilt model", *J. Acoust. Soc. Am.,* 107(3):1697-1714, 2000.

[17] Möhler, G., "Describing intonation with a parametric model", in *Proc. ICSLP, Sydney 1998*, pp. 2851-2854

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

[19] Loh, W.Y. and Shih, Y.S., "Split selection methods for classification trees", Statistica Sinica (7):815-840, 1997.

[20] Statistica 6.0: Statistica for Windows [Computer program] StatSoft, Inc. (2001)

[21] Breuer, S., Stober, K., Wagner, P. and Abresch, J., „Dokumentation zum Bonn Open Synthesis System BOSS II". Unpublished document, IKP, Bonn (2000)

[22] Demenko, G., Wypych, M. and Baranowska, E., "Implementation of Polish grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis", *Speech and Language Technology,* (7):79-96, 2003.

[23] Demenko, G. and Wagner A., "Prosody annotation for unit selection text-to-speech synthesis", *Archives of acoustics,* 32(1):25-40, 2007.

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

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

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

[27] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P,. "Segmental durations in the vicinity of prosodic phrase boundaries", *J. Acoust. Soc. Am.,* 91(3):1707-1717, 2000.

[28] Horne, M., Strangert, E. and Heldner, M., "Prosodic boundary strength in Swedish: final lengthening and silent interval duration", in *Proc. 13th Int. Cong. Phon. Sci., Stockholm 1995,* pp. 170-173

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

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

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

[32] t'Hart, J., Collier, R. and Cohen, A., *A Perceptual Study of Intonation,* Cambridge Uni. Press, Cambridge, 1990.

[33] Hirschberg, J., "Communication of prosody - functional aspects of prosody", *Speech Comm.,* 36(1):31-43, 2002.

[34] Syrdal, A. Möhler, G., Dusterhoff, K., Conkie, A. and Black, A. W., "Three Methods of Intonation Modeling", in *Proc. 3rd SSW, 1998,* pp.305-310