INTERSPEECH 2009 BRIGHTON

Analysis and Recognition of Accentual Patterns

Agnieszka Wagner

Department of Phonetics Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland

wagner@amu.edu.pl

Abstract

This study proposes a framework of automatic analysis and recognition of accentual patterns. In the first place we present the results of analyses which aimed at identification of acoustic cues signaling prominent syllables and different pitch accent types distinguished at the surface-phonological level. The resulting representation provides a framework of analysis of accentual patterns at the acoustic-phonetic level. The representation is *compact* - it consists of 13 acoustic features, has *low redundancy* – the features can not be derived from one another and *wide coverage* – it encodes distinctions between perceptually different utterances. Next, we train statistical models to automatically determine accentual patterns of utterances using the acoustic-phonetic representation which involves two steps: detection of accentual prominence and assigning pitch accent types to prominent syllables. The *efficiency* of the best models consists in achieving *high accuracy* (above 80% on average) using small acoustic feature vectors.

Index Terms: prominence, pitch accents, prosody labeling

10.21437/Interspeech.2009-306

1. Introduction

Every spoken utterance is characterized by specific segmental structure and prosody – rhythm, intonation and stress pattern. These aspects of utterances are realized to a different extent by variation in the prosodic parameters: F0 (fundamental frequency), intensity, duration and voice quality. Among the most important communicative functions of prosody there are *prominence* and *phrasing*. Variation in prominence and/or phrasing changes the meaning hearers assign to an utterance. In the current study we concentrate on prominence and propose a framework of automatic analysis and recognition of accentual patterns.

1.1. Prominence and accentual patterns

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 that a word or its part stands out from its environment [1]. Prominence is closely related to *information structure*. First of all, in some languages accentuation of words serves to distinguish between *new* (accented) and *old* or *given* (unaccented) items. Secondly, different parts of an utterance receive prominence of a different *degree* in order to highlight those parts which deserve special attention. The most salient, prominent part is known as *focus*. Apart from the binary opposition accented vs. unaccented distinctions are drawn among various types of pitch accents. The latter 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 [2]. Thus, the use of a specific accentual pattern determined by accenting some words and failing to accent others together with realization of specific pitch accent types serves communication purposes.

1.2. Automatic recognition of prosodic events

The fact that in the communication process human listeners rely heavily on prosodic cues has led to a growing interest in incorporating prosodic information in various speech applications. The use of prosodic information can improve the performance of speech recognition systems and is crucial for generating a naturally sounding speech by text-to-speech systems. Despite criticism concerning various aspects of higher-level representations of prosody such as ToBI or INTSINT they can be very useful in speech technology: they encode both melodic and functional aspects of prosody and provide information which helps to learn relationships between acoustic-prosodic parameters and segmental acoustics on the one hand and lexical and syntactic structure of utterances on the other.

In order to effectively use the information conveyed by prosody a number of approaches has been proposed which enable automatic determination (in speech synthesis scenario) or labeling (in speech understanding or recognition scenario) of an accentual pattern of a given utterance. Most of the existing models performing automatic detection and labeling of accentual patterns rely on large vectors of features (up to 276 features in [3]) of a different type (e.g., acoustic, lexical, syntactic) which are considered as the main cues signaling accentual prominence and discriminating well between different pitch accent categories (e.g., [4], [5]).

The statistical modeling techniques applied to automatic labeling of accentual patterns include neural networks [6], classification trees [7], maximum entropy models [8], discriminant function analysis [9] or HMMs. The models' complexity depends on speaking style (neutral, reporting style vs. spontaneous or conversational speech), the number of categories 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 e.g. identification of prominent syllables and labeling of pitch accent types.

The best models performing detection of accentual prominence achieve an overall accuracy between 70% [5] and 87% [4]. In manual annotation of pitch accent position the inter-transcriber agreement is 80%-90% (e.g., [10], [11]).

In the automatic classification of pitch accent types the best models yield accuracy of about 80% ([4], [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 [10] and almost 95% in [11]

1.3. Features of the current approach

The objective of this study is to provide a framework of automatic analysis and labeling of accentual patterns. To that end we investigate the acoustic-phonetic realization of accentual prominence and different pitch accent types distinguished at the surface-phonological level (see sec. 3.2).

As a result two small vectors of acoustic features are defined. They 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 (vowel/syllable type, phone/syllable/word boundary, lexical stress) and exclude intensity features (cf. [3], [4], [8]). With few exceptions our acoustic features refer to relative values i.e., values normalized against variation due to prosodic structure (in case of F0 features, see [12]) or caused by intrinsic properties of phonemes and especially vowels (in case of duration features).

The resulting representation provides a framework of analysis of accentual patterns at the acoustic-phonetic level. The representation is compact, has low redundancy and wide coverage. From this representation a higher-level surface-phonological description of accentual patterns can be easily and reliably derived. The latter is in terms of discrete tonal categories of pitch accents, and contrary to the acoustic-phonetic representation, it encodes not only melodic, but also functional aspects of prosody.

We apply statistical modeling techniques such as neural networks (NN), classification trees and discriminant function analysis (DFA) to design models performing detection and classification of pitch accents. Unlike ([6], [8]) where syntactic features are also used, the models designed in the current study rely only on acoustic feature vectors. Contrary to other authors e.g., [5] we design two different models and train them with different feature sets to perform one of the two tasks: accentual prominence detection or pitch accent type classification. Prominence detection is performed at the word level i.e., only lexically stressed syllables are taken into account. 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 a different complexity and trained with different prior classification probabilities are tested, and the most efficient ones are selected for further optimization. As regards classification trees the QUEST classification tree program is used. The models are designed semi-automatically using Statistica 6.0.

2. Speech material and feature extraction

2.1. Speech corpus

The speech material used in the current study comes from the corpus used in the Polish module of BOSS unit selection TTS system. The corpus contains recordings of phonetically rich and balanced sentences, fragments of fiction and reportage read in a neutral, reporting style.

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. Manual annotation consisted in identification of the position and types of pitch accents and phrase boundaries. As shown in a number of studies ([10], [11]) there are substantial discrepancies in the annotation of prosody provided manually by different labelers. Therefore, in order to minimize the inconsistency in the labeling of pitch accent types we use data labeled by a single annotator. Consequently, the subset of the unit selection speech corpus used in the current study consists of 1052 utterances (15566 syllables including 6417/3926 stressed/pitch accented syllables) representative of the whole speech material included in the corpus.

2.2. Data preparation for analyses

2.2.1. F0 extraction and processing

F0 extraction and processing was performed with a Praat (ver. 4.1.5) script. F0 was extracted every 10 ms (based on 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 (5Hz bandwidth). The unvoiced regions were interpolated through. The waveforms were resynthesized with the smoothed and interpolated F0 contours using PSOLA. The visual and perceptual inspection confirmed that the resulting F0 data can be regarded as a reliable basis for parameterization and analysis.

2.2.2. Feature extraction

For the analysis of the acoustic-phonetic realization of pitch accents for each syllable and its vocalic nucleus a number of features describing variation in the fundamental frequency and duration was automatically extracted with a Praat script. Some of the features refer to absolute values while others to relative values i.e. values normalized with respect to overall F0 level over the length of the phrase (in case of F0 parameters e.g., *mean, maximum and minimum F0*) or expected duration determined for a given vowel or syllable type (in case of duration parameters: *syllable and nucleus duration*). Other features describe pitch variation in a two-syllable window including the current and next syllable or alternatively, the nucleus of the current syllable and that of the previous syllable. For each syllable and its vocalic nucleus features of the two previous and two next syllables and vowels were provided as well. The resulting inventory (for details see [12]) consists of acoustic features which are commonly used in the analysis of intonation e.g., [13].

3. Analysis of accentual patterns

3.1. Acoustic correlates of accentual prominence

Generally, there is no agreement as regards acoustic correlates of stress and prominence. In some studies e.g., [1] *variation in fundamental frequency* (or pitch movements) and *overall intensity* are identified as the main cues signaling accentual prominence, while *duration* and *spectral emphasis* are said to play a secondary role in this respect [14]. In the current study we rely most of all on F0 features, which is in accordance with the results presented in [15] which proved that variation in the fundamental frequency is the main correlate of accentual prominence in Polish.

In a series of ANOVA and DFA analyses (details are given in [12]) we looked for statistically significant differences in the acoustic-phonetic realization between unaccented and accented syllables in terms of the acoustic parameters listed in sec. 2.2.2 and examined correlations between them in order to avoid redundancy. Consequently, we identified five features describing variation in F0 and duration 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 [4]
- Tilt (F=648.47) – a parameter describing the shape of the pitch movement on the syllable calculated as in [13]
- height of F0 peak on the syllable (F=591.4)

The results of our analyses showed that prominent syllables are characterized by almost twice as high average slope as unaccented syllables (131.18 vs. 73.39 Hz/s) which indicates that 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) – the latter shows that most pitch accents are realized by both rising and falling pitch movement and that the rising pitch movements are generally more relevant to prominence then pitch falls. Accented syllables/vowels are much longer than the stressed unaccented ones - the average difference is 20ms/10ms. The effects discussed here are statistically significant ($p<0.01$).

The features: slope, F0 peak, Tilt, relative syllable and nucleus duration can be easily derived from utterance's acoustics and lexical features. With one exception (i.e. relative syllable and nucleus duration) they are not significantly correlated with one another, which ensures *low redundancy* of the acoustic-phonetic representation of accentual prominence.

3.2. The inventory of pitch accents at the surface-phonological level

The surface-phonological description of accentual patterns used in the current study was proposed in [12]. It was derived from the prosody labeling framework applied in the Polish unit selection corpus (see sec. 3.1.2) 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 marks a lower and H a higher tonal target, and the asterisk indicates which of the two tones is aligned with the accented vowel.

This representation encodes both melodic and functional aspects of prosody. From strictly phonological representations it differs in that it makes no distinction between linguistic and paralinguistic functions of prosody. It is assumed that surface-phonological descriptions can be more useful for application in speech technology than strictly phonological systems, because they are more perceptually-oriented.

3.3. Acoustic-phonetic realization of different pitch accent types

In this section we report on the results of the analyses of the acoustic-phonetic realization of different pitch accent types. In order to identify features which discriminate the best among different categories the effect of pitch accent type on variation in the parameters derived from utterance's acoustics and lexical features (see sec. 2.2.2) was investigated in a series of ANOVA and DFA analyses. Correlation matrices showing the relations between the parameters were examined too in order to avoid redundancy. As a result, a vector consisting of eight features was defined. It includes:

- amplitude of the rising (F=631.9) and falling F0 movement (F=1537.9) determined for the vowel
- relative mean (F=591.4), maximum (F=198.2) and minimum F0 (F=551) determined for the vocalic nucleus
- Tilt (F=1258.4) and Tilt amplitude (F=1042.2)
- direction calculated as a ratio of mean F0 (F=650.6) and like Tilt and Tilt amplitude determined in a two-syllable window containing accented and the next syllable

The effect of pitch accent type on variation in the features listed above is statistically significant ($p<0.01$). The resulting representation is compact and has low redundancy (except for Tilt and Tilt amplitude no significant correlations among the features were found). Together with the features identified as the acoustic correlates of prominence (sec. 3.1) it constitutes the acoustic-phonetic description of accentual patterns.

4. Recognition of accentual patterns

4.1. Automatic detection of prominent syllables

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 4278 and 2139 stressed syllables respectively. The ratio of accented (+acc) to unaccented (-acc) syllables was about 2:1 in each sample. The models were trained using the vector of 5 acoustic features described in sec. 3.1. The best discrimination between accented and unaccented syllables was achieved with neural networks: in the test sample the RBF 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 DFA: 79.13% (test sample) and 77.23% (cross-validation test) respectively. The table below summarizes the average accuracy of correct detections of accented and unaccented syllables in the test sample (or cross-validation test in case od DFA). The numbers in brackets (column *class)* show chance-level accuracies.

Table 1: *Detection of prominence at the word level - summary of models' performance (test sample).*

class	MLP (5:17:1)	RBF (5:82:1)	class. 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**

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 ([3], [7], [15]) and to the levels of agreement between human labelers e.g., 80.6% in [10]. The best-performing models yield between 84-87% accuracy (e.g., [4], [5], [6], [8]), but the advantage of our approach is the use of a compact and simple representation consisting of five features 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 the findings presented in the literature e.g., [15] 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.

4.2. Classification of pitch accent types

The models presented in this section derive automatically the surface-phonological description of pitch accents from the acoustic-phonetic description defined in terms of two vectors of acoustic features describing the realization of accentual prominence and different pitch accent types. The models (DFA, NN and classification trees) were trained on a subset of 3671 syllables marked as being associated with a pitch accent. Syllables associated with different pitch accent types were

proportionally divided into a training (2754) and test sample (917). The models rely on a vector consisting of 8 acoustic features which can be easily derived from utterance's F0, timing cues and lexical features (see sec. 3.3).

The performance of the classification tree (28 splits and 29 terminal nodes) was superior to the that of discriminant function analysis and neural networks. In the test sample the classification tree outperformed other models achieving an overall accuracy of 81.67% which compares favorably with ([7], [16]). Generally, the highest overall recognition rate was achieved for HL* accents, whereas the recognition of LH*L accents was the most problematic one. Table 2 summarizes the average accuracy of correct detections of pitch accent types in the test sample. It can be seen that the results are much better than a chance-level pitch accent type assignment (chance-level accuracy is given in brackets, column class).

Table 2: *Pitch accent type classification – summary of models' performance (test sample).*

class	MLP (8:15:5)	RBF (8:82:5)	class. tree	DFA
H*L (36.86)	76.63	78.99	71.01	71.89
L*H (10.25)	77.66	70.21	86.17	70.21
LH* (28.9)	83.02	85.66	70.19	80.38
HL* (20.94)	89.58	89.58	91.67	88.54
LH*L (3.05)	60.71	39.29	89.29	85.71
Average (%):	**77.52**	**72.75**	**81.67**	**79.35**

As regards contribution of particular acoustic features into the recognition accuracy the importance ranking of predictor variables generated for the classification tree showed that features describing the shape and amplitude of the pitch movement were the most important, whereas relative peak and minimum height were the least important.

The advantage of our approach is high average accuracy which exceeds the levels of agreement among human labelers in manual pitch accent type assignment reported in some studies e.g., [10] and the use of a simple and compact acoustic-phonetic representation.

5. Discussion and conclusions

The objective of the current study was to propose a framework of an efficient automatic analysis and recognition of accentual patterns. For that purpose two feature vectors were defined which constitute representation of the acoustic-phonetic realization of accentual prominence and different pitch accent types distinguished at the surface-phonological level. The acoustic-phonetic representation is *compact* - it consists of 13 features and *simple* - it can be easily derived from utterance's acoustics (F0, timing cues) and lexical features. The representation has *low redundancy* – its components can not be derived from one another and *wide-coverage* – it encodes distinctions between perceptually different utterances.

The advantage of our framework of automatic recognition of accentual patterns 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. The efficiency of the best models is ensured by the fact that they yield accuracy comparable to that reported in other studies and approaching the levels of agreement among human annotators in manual labeling of accentual patterns.

Work is in progress to design models performing a speaker-independent automatic labeling of accentual patterns based on the framework proposed in the current study. In the future, different speaking styles will be taken into account too.

6. Acknowledgements

The research presented in this paper was supported by The Polish Scientific Committee (Project ID: R00 035 02).

7. References

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

[2] Ladd, R., "Intonation", Cambridge University Press, Cambridge, 1996.

[3] 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

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

[5] 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

[6] 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.

[7] 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

[8] 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.

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

[10] 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

[11] 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

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

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

[14] 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

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

[16] 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

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