Automatic labeling of prosody

Agnieszka Wagner
Department of Phonetics, Adam Mickiewicz University, Poland
https://doi.org/10.36505/ExLing-2008/02/0056/000115

Abstract

The paper proposes a framework for automatic prosody labeling. The labeling involves detection of the location of accented syllables and phrase boundaries, and recognition of pitch accent and boundary tone types. A number of classification models are designed to perform these tasks on the basis of small vectors of acoustic features. The models achieve high accuracy and their performance is comparable to the results reported in other studies on automatic prosody labeling and to inter-labeler consistency in manual labeling of prosody.

Key words: prosody, description, automatic recognition/detection

Introduction

In recent years a growing interest in automatic labeling of prosody is observed, which can be attributed to the development of speech applications such as speech synthesis or recognition. Generally, all types of speech applications require speech corpora which have to be provided with appropriate annotation. Manual annotation of prosody is laborious, time-consuming and not entirely consistent e.g., Pitrelli et al. (1994), Grice et al. (1996), Yoon et al. (2004). A solution to these problems is development of models providing the labeling of prosody automatically e.g., Wightman and Ostendorf (1994), Kiessling et al. (1996), Sridhar et al. (2007).

In the next sections methods capable of identifying the prosodic structure of utterances and providing the description of prosodic events (pitch accents, boundary tones) on the surface-phonological level are proposed. They were designed on the basis of a subset of the unit selection speech corpus used in the Polish module for BOSS, Breuer et al. (2000). The speech material consisted of 1052 utterances read by a professional male speaker in a news-reading style, but examples of expressive speech were also provided. The corpus was automatically segmented on the phoneme/syllable/word level; Stress was assigned from rules. Prosodic labeling was done manually.

Decision trees and neural networks (MLP, RBF and linear networks) were applied to solve the detection/recognition problems. In the next sections only the results achieved by the best models are reported. All the models were designed using the *Statistica Neural Network* package available in Statistica 6.0 (2001).

Automatic labeling of boundary tones

It involves two steps, namely detection of phrase boundary location and recognition of boundary tone type. The former is performed on the word-level i.e. only word-final syllables are taken into account, which ensures that only one boundary per word can be identified. The latter is performed only for phrase-final syllables. The inventory of boundary tones consists of 2 rising (labeled 2,? and 5,?) and 3 falling boundaries (2,. 5,. and 5,!). They are distinguished on the basis of direction and amplitude of the distinctive pitch movement and scaling of f0 targets at the start and end of the movement. Boundary tones labeled 2,? and 2,. are associated with minor phrase boundaries, whereas 5,. 5,? and 5,! – with major phrase boundaries. The inventory is part of the prosody description on the surface-phonological level which encodes both melodic and functional aspects of prosody. As shown in Wagner (2008) this description provides information which is highly significant to the estimation of pitch target level for F0 generation in speech synthesis.

Detection of phrase boundary location

For this task the following acoustic features were used: 1) relative duration of the nucleus of the current and previous syllable, 2) relative duration of the current syllable, 3) F0 features determined for the vocalic nucleus - tilt value, rising amplitude, overall F0 level and slope. The tilt and amplitude parameters were calculated as in Taylor (2000). Relative durations were calculated as in Rapp (1998) and were used to eliminate the effect of syllable structure and/or vowel type on the observed duration.

The highest accuracy of phrase boundary detection was achieved with a radial basis function (RBF) network with 23 neurons in the hidden layer. The network was trained using 5844 syllables and tested on a subset consisting of 1000 syllables. In the test sample the location of phrase boundary was identified with **81,6%** accuracy, whereas syllables of a non-final position in the phrase were correctly identified in **79,3%**.

Recognition of boundary tone type

The vector of acoustic features used in this task consisted of: 1) syllable-final F0 value, 2) overall F0 level on the previous syllabic nucleus, 3) direction of the pitch, and 4) distance to the next silent pause measured in the number of syllables.

The RBF network with 54 neurons in the hidden layer had the best performance – it achieved an overall accuracy of **87,6%** (test sample). The network was trained on a subset of 1132 syllables and tested on 377 syllables. The boundary tones labeled 5,. were identified with the highest

accuracy i.e. 98,6%, whereas boundaries labeled 2,. were correctly recognized in 70,13%.

Automatic labeling of pitch accents

The first task involves detection of accented syllable position and it is performed on the word-level i.e. only stressed syllables are taken into account. In this way only one accented syllable per word can be identified. Then, the types of pitch accents distinguished on the surface-phonological level are recognized. In this latter task only accented syllables are taken into account. The inventory of pitch accents includes 2 rising accents (labeled L*H and LH*), 2 falling accents (H*L and HL*) and 1 rising-falling accent (LH*L). They are distinguished on the basis of the direction of the pitch movement, timing of the F0 peak relative to the accented syllable onset and range of an f0 change on the vocalic nucleus. Together with the inventory of boundary tones, the pitch accent inventory constitute the description of prosody on the surface-phonological level.

Detection of accented syllable location

The detection of accented syllable position was based on 2 duration features including relative duration of the syllabic nucleus and syllable, and 3 F0 features describing the amount of pitch variation on the syllable, peak height and tilt value of the syllable. Altogether 6417 stressed syllables were used in the experiments with 3929 accented syllables among them.

MLP and RBF networks performed significantly better then the decision tree or the linear network. The MLP network achieved **81,65%** accuracy in the detection of accented syllables and **81,79%** in the detection of unaccented syllables (test sample). The results for the RBF network were **82,14%** and **81,76%** respectively.

Recognition of pitch accent type

The recognition of pitch accent types required a larger acoustic feature vector consisting of parameters describing the amplitude of the pitch movement on the vowel, direction of the pitch movement (calculated as a difference in mean F0 on the vowel between the current and next syllable), tilt value determined in a 2-syllable window containing the current and next syllable, F0 peak, minimum and mean associated with the accent and normalized relative to the F0 mean in the phrase.

The best results were achieved with a classification tree including 27 splits and 28 terminal nodes. The tree was designed using QUEST classification programme available in Statistica 6.0. 2754 syllables were used for training and 917 for testing. The tree performed with an overall

accuracy of **81,63%** (test sample). LH*L accents were recognized with the highest accuracy i.e. 89,3%, whereas LH* accents were correctly identified in 70,2%.

Discussion and conclusions

The models presented in this paper recognize the prosodic structure of utterances and provide the surface-phonological description of prosody in terms of different types of pitch accents and boundary tones.

The performance of the models is comparable to the results reported in other studies on automatic prosody labeling e.g. Rapp (1998), Sridhar et al. (2007) and inter-labeler consistency in manual transcription of prosody e.g. Grice et al. (1996). The advantage of the models proposed in this paper is that they require only small vectors of acoustic features (as oppossed to e.g. 276 features used in Kiessling et al. (1996)) which can be easily derived from utterance's acoustics.

References

Breuer S., Stober K., Wagner P. and Abresch J. 2000. Dokumentation zum Bonn Open Synthesis System BOSS II. Project report, IKP, University of Bonn.

Grice M., Reyelt M., Benzmueller R. and Mayer J., Batliner A. 1996. Consistency in transcription and labeling of German intonation with GToBI. Proc. of the 4th ICSLP, 1716-1719, Philadelphia, USA.

Kiessling A., Kompe R., Batliner A., Niemann H., Nöth E. 1996. Classification of Boundaries and Accents in Spontaneous Speech. Verbmobil-Report 156, University of Erlangen-Nuremberg, University of Munich, August 1996.

Pitrelli J.F., Beckman M.E. and Hirschberg J. 1994. Evaluation of prosody transcription labeling reliability in the ToBI framework. Proc. of the 3rd ICSLP, 123-126, Yokohama, Japan.

Rapp S. 1998. Automatic Labeling of German Prosody. Proc. of the 5th ICSLP, Sydney, Australia.

Sridhar V. K. R., Bangalore S., Narayanan S. 2007. Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework. IEEE Transactions on Audio, Speech and Language Processing, 16(4), 797-811.

Statistica 6.0 2001. Statistica for Windows [computer program], StatSoft, Inc., Tulsa

Taylor P. 2000. Analysis and synthesis of intonation using the Tilt model. J. of Acoust. Soc. Am. 107(3), 1697-1713.

Wagner A. 2008. A comprehensive model of intonation for application in speech synthesis. PhD thesis, Institute of Linguistics, Adam Mickiewicz University.

Wightman C. and Ostendorf M. 1994. Automatic labeling of prosodic patterns. IEEE Transactions on Audio, Speech and Language Processing, 2(4), 469-481.

Yoon T. J., Chavarría S., Cole J. and Hasegawa-Johnson M. 2004. Intertranscriber Reliability of Prosodic Labeling of Telephone Conversation Using ToBI. Proc. of the 8th ICSLP, 2729-2732, Jeju Island, Korea.