INTERSPEECH 2009 BRIGHTON

Experiments on Automatic Prosodic Labeling

Antje Schweitzer[1], *Bernd Möbius*[1,2]

[1]Institute of Natural Language Processing, University of Stuttgart, Germany
[2]Institute of Communication Sciences, University of Bonn, Germany
antje.schweitzer@ims.uni-stuttgart.de, bernd.moebius@ims.uni-stuttgart.de

Abstract

This paper presents results from experiments on automatic prosodic labeling. Using the WEKA machine learning software [1], classifiers were trained to determine for each syllable in a speech database of a male speaker its pitch accent and its boundary tone. Pitch accents and boundaries are according to the GToBI(S) dialect, with slight modifications. Classification was based on 35 attributes involving PaIntE F0 parametrization [2] and normalized phone durations, but also some phonological information as well as higher-linguistic information. Several classification algorithms yield results of approx. 78% accuracy on the word level for pitch accents, and approx. 88% accuracy on the word level for phrase boundaries, which compare very well to results of other studies. The classifiers generalize to similar data of a female speaker in that they perform equally well as classifiers trained directly on the female data.
Index Terms: perception of prosody, prosodic labeling, F0 parametrization

1. Introduction

The following research question is at the bottom of the experiments presented in this paper: What are the targets in prosody production? In an exemplar-theoretic view, the targets in prosody production are derived from exemplars stored in memory, and they are used in perception to categorize new exemplars. The aim of this paper is to model a listener's memory by a large prosodically annotated speech database and to simulate prosodic categorization using machine learning methods to classify new exemplars. Successful simulation of prosodic categorization is not only a step towards understanding perception, it also has an application in automatic prosodic labeling, which is known to be very time-consuming.

We used two speech databases that had been annotated on the segment, syllable, and word level, and prosodically labeled according to GToBI(S) [3]. Prosodic labeling had been carried out in the earlier SmartWeb project [4] without having automatic prosodic labeling in mind.

The databases were converted to the Festival [5] utterance format, and the F0 contour of each utterance was PaIntE parametrized [2]. Finally, for each syllable, all properties that were potentially relevant in classification were derived from the Festival utterance structures and captured in 37 attributes. Thus, the speech databases are represented as sets of instances of syllables. Each syllable instance is characterized by the 37 attributes. Two of these attributes represent the type of pitch accent and the type of boundary tone realized on the syllable.

Using the first database as training data, we applied various machine learning schemes implemented in the WEKA software [1] to build classifiers for both syllable-based accent and boundary prediction, i.e. to build classifiers that decide on the value of the **accent** or **boundary tone** attribute of a syllable instance based on the values observed for the remaining attributes. In order to compare the results of the present study to studies which just predict two classes of accent (no accent vs. accented) and two classes of boundaries (boundary vs. no boundary), classifiers for these two-class problems were trained in addition to the classifiers predicting the full set of pitch accents and boundaries. Except for one case (for IBk instance-based learning), the default parameters suggested by WEKA were used in building the classfiers[1].

The performance of the learning algorithms was compared, and the classifiers built according to the best learning algorithm were applied to the second, very similar database of another speaker in order to assess the generalizability of the classifiers.

2. Experiments

2.1. Data for training and testing the classifiers

The speech databases are the SWMS database (2 hrs., male) and the SWRK database (3 hrs., female), which have been recorded in the course of the SmartWeb project [4]. The speakers are professional speakers of Standard German. For both databases, the utterances represent typical utterances of 5 different genres. They were read off a screen at recording time.

The databases were originally used for unit selection speech synthesis in the SmartWeb project. The recording procedure and the prompts were identical for both databases, but there are more utterances in the SWRK database than in the SWMS database. Both databases were split into a training and a test set with the test set consisting of about 10% of the utterances of the original databases. The splits were identical for both databases. The SWMS test set was not used for building the classifiers at all because its utterances are also contained in the test set of the SWRK data, on which the classifiers were to be evaluated later when assessing their generalizability.

2.2. PaIntE parametrization

Several of the attributes used for classification involve parameters obtained by PaIntE [2]. PaIntE stands for "Parametrized Intonation Events" and was originally intended for F0 modeling in speech synthesis. The basic idea was to approximate the F0 contour in a certain window around syllables that are known to carry a pitch accent or a boundary tone using a linguistically motivated approximation function. A schematic of the function is given in figure 1. It is composed of a rising and a falling

[1]The default settings of the IBk learning scheme implemented in WEKA set the k parameter to 1. The k parameter determines the number of neighbours considered in classification of new instances, and with k=1 this learning scheme is identical to the separate IB1 learning scheme in WEKA. Therefore, k=30 was used instead.

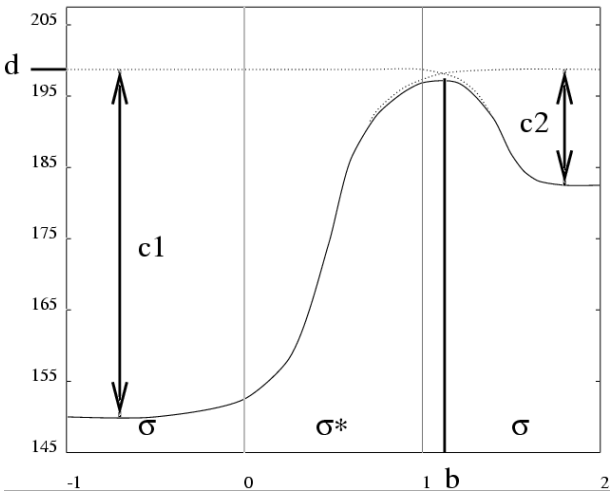

Figure 1: *Schematic of the PaIntE approximation function, reproduced from [2]. The approximation window represents three syllables. The accented syllable is indicated by the asterisk (σ*). Peak height is determined by parameter d, amplitudes of rise and fall correspond to parameters c1 and c2, respectively, and peak alignment depends on the b parameter.*

sigmoid function. The exact contour is determined by six parameters a1, a2, b, c1, c2, and d, where a1 and a2 represent the (amplitude-normalized) steepness of the rising and falling sigmoid, respectively, and c1 and c2 specify the amplitudes of the sigmoids. Parameter d can be interpreted as approximating the absolute peak height in Hertz, and parameter b determines the alignment of the peak in terms of relative position in the normalized duration of the three syllables.

Usually, the size of the approximation window is influenced by prosodic structure: the window is not extended beyond phrase boundaries. Also, parametrization is carried out for syllables known to be accented only. In contrast, we have applied the parametrization to every syllable, always using a three-syllable window. To this end, the PaIntE source code has been modified to allow for parametrization without any assumptions about, or references to, prosodic properties derived from the prosodic labels.

2.3. Attributes used in classification

The **accent** attribute indicates which, if any, pitch accent has been realized on the syllable. Accents are according to the GToBI(S) [3] labeling system. The accent attribute can assume the following values: NONE for unaccented syllables, or L*H, H*L, L*, H*, ..H, ..L, L*HL, or HH*L. For the **boundary tone** attribute, we have mapped the underspecified boundary tones % and - of the GToBI(S) system to fully specified tones by integrating the preceding trail tone into the boundary tone labels. To distinguish them from boundary tones that had been fully specified already, the preceding trail tones are specified in brackets. The full set of possible values for the tone attribute is NONE for non-final syllables, and (H)-, (L)-, (H)%, H%, (L)%, or L% for phrase-final syllables. The values of the accent and the boundary tone attributes are to be predicted by automatic classification based on the values of the remaining attributes.

Attributes **a1**, **a2**, **b**, **c1**, **c2**, and **d** correspond to the six PaIntE parameters. To eliminate speaker-specific aspects, all parameters but the b parameter were z-scored. The b parameter was left unchanged because we did not expect speaker-specific effects for this parameter. Further attributes are derived from the PaIntE parameters: **maxc** is the maximum of c1 and c2, and **c1-c2** codes the relative difference in F0 before and after the F0 movement by subtracting the amplitude of the falling movement, c2, from the amplitude of the rising movement, c1. PaIntE parameters c1, c2 and d of the preceding two and of the following syllables are also taken into account.

Turning to temporal aspects, we have claimed that phoneme-specific z-scores of segment durations are relevant in the perception of prosodic events[2] [6]. Both the z-scores for nuclei (**zscnucleus**) and the z-scores for final segments (**zscfinalseg**) are provided as attributes. In order to facilitate comparison of values for the current syllable with those for context syllables, these attributes are provided for a three-syllable window around the current syllable.

The **stress** attribute indicates whether a syllable is stressed or not, **wordfin** indicates whether the syllable is word-final or not, and **silnext** specifies whether a silence follows. The remaining attributes are "text-based" attributes which are derived from orthographic information and punctuation marks. These attributes are a subset of the attributes used by the prosody prediction module of the IMS Festival text-to-speech synthesis system [7]. Thus, they are known to be predictive of prosodic structure. Attribute **pos** specifies the part-of-speech (POS) tag of the word that the syllable is related to. POS tags were obtained by the German Tree Tagger [8]. We have also included the attribute **func**, which maps the POS tags to the two classes function and content word. The attribute **top** is also derived from the POS tags. It specifies whether the syllable is at the end of the so-called "Vorfeld". Another concept that is used for predicting prosodic events in our TTS system is the noun chunk [9]. We use the POS tags to identify noun chunk boundaries. Final nouns in noun chunks are interpreted as head of the noun chunk (attribute **head**). The weight of the chunk in terms of number of content words is also used (**wght**). The last attribute, **punc**, specifies whether there was a punctuation symbol after the syllable in the text underlying the utterance (i.e., in the text the speaker was prompted with).

3. Results

Performance was measured in terms of prediction accuracy. The accuracy rate is the proportion of instances which is correctly classified by a classifier. In order to assess performance of the classifiers on the word level we have, for accent classification, eliminated all lexically unstressed syllable instances and kept only the one lexically stressed syllable per word on which pitch accents would have to be realized; analogously, for boundary classification, we have eliminated all instances of word-internal syllables. Thus, in both cases, only the one relevant syllable of the corresponding word is classified, and the accuracy rates obtained on the transformed data sets can be interpreted as word-based accuracies. This allows for comparing the present results

[2]Phoneme-specific z-scores z(dur_i) of phone durations are obtained by subtracting from the specific duration dur_i the mean of all observed durations of realizations of the same phoneme type, mean(dur_P), and dividing by the standard deviation of durations of the same phoneme type, sd(dur_P):

$$z(dur_i) = \frac{dur_i - mean(dur_P)}{sd(dur_P)} \quad (1)$$

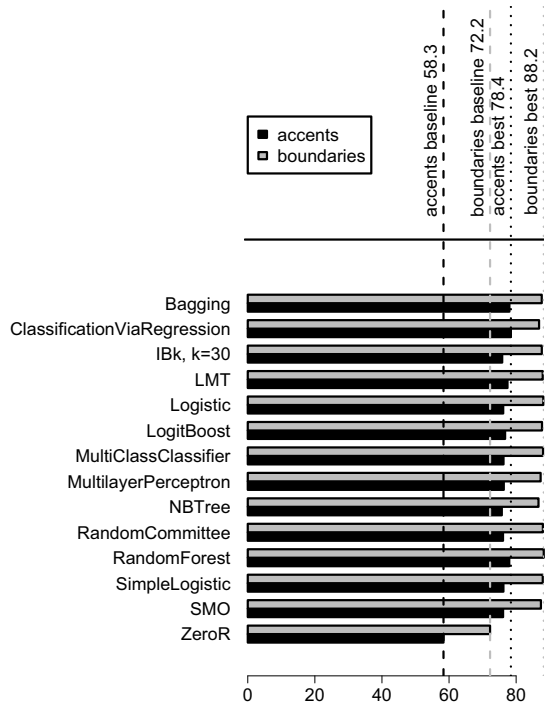

Figure 2: *Word-based averaged accuracy rates of those machine learning algorithms that yielded the best results. The bars indicate accuracy rates for accent classification (black bars) and boundary classification (grey bars). The ZeroR algorithm was included as a baseline.*

to many other studies which only report on word-based measures of performance.

To compare the suitability of various learning algorithms for the present problem, the accuracy of the resulting classifiers was estimated in three separate runs using 10-fold stratified cross-validation in each run. Thus, the accuracy rates correspond to averaged accuracy rates of 30 different classifiers built from various splits of the training data. Separate classifiers were trained for classifying accents and boundaries, and for the two-class vs. the original class problem, thus, for each algorithm, 120 classifiers were built altogether.

Figure 2 is intended to give an impression of the performance of the learning schemes that yielded the best results for the present experiments. It presents word-based averaged accuracy rates of all classification schemes implemented in WEKA (version 3.4) for which we observed rates better than 75% for accent classification and better than 86% for boundary tone classification. The accuracy rates of the classifiers for accent classification are indicated by black bars, those of the classifiers for boundary classification are indicated by grey bars. The classifiers are listed in alphabetical order. To avoid confusion, we have kept the original names of the algorithms from the WEKA implementation.

In figure 2, the vertical dashed lines represent the baselines, which are determined as the word-based accuracy rates achieved by the ZeroR learning algorithm, which are indicated at the very bottom of figure 2. The ZeroR classifier just assigns

Table 1: *Word-based averaged accuracy rates for the best algorithms and the baseline. Most accuracy rates are comparable to the rates obtained by the RandomForest algorithm: only accuracy rates marked by * are statistically significantly worse than the corresponding rate for the RandomForest algorithm. Class.Regression is abbreviated for WEKA's ClassificationViaRegression.*

Algorithm	accents		boundaries	
	2-class	full set	2-class	full set
Bagging	86.19	78.08	93.33	88.00
Class.Regression	85.49	78.17	*92.29	*87.41
LMT	86.24	77.54	93.37	87.84
RandomForest	86.17	78.04	93.31	88.16
ZeroR	*58.30	*58.23	*72.16	*72.16

all instances the most frequent class found in the training set and thus can be interpreted as providing the chance level.

The two dotted lines indicate the best results obtained in terms of averaged word-based accuracy - the black dotted line shows the best rate in accent classification, which was at 78.2%, and which was obtained using the ClassificationViaRegression scheme; the grey dotted line indicates the best rate for boundary tone classification, which was at 88.2% and was obtained by the RandomForest scheme. Figure 2 illustrates that the best results are not due to one or few outstanding learning algorithms that are particularly suitable for the present data; rather, when providing the information coded in the attributes discussed above, one can reliably reach quite high accuracy rates using various learning algorithms.

The exact word-based averaged accuracy rates of the best learning algorithms from figure 2 are listed in table 1. Accuracy rates are given both for the two-class problem and for predicting the full set of GToBI(S) lables. To assess whether any of the algorithms is better than the rest, the asterisks indicate which classifiers performed significantly worse than the RandomForest classifiers. It can be seen that several classifiers perform equally well, without significant differences in the accuracy rates.

3.1. Generalizability

As a first step towards assessing the generalizability of the classifiers to other data, the best classifiers have been applied to the test set of the (female) SWRK database. The results are comparable for boundaries: for instance, a RandomForest classifier trained on the full SWMS training data reaches accuracy rates of 88.9% on the SWMS test data and of 88.6% on the SWRK test data.

For pitch accents, performance is lower on the SWRK data: for instance, a RandomForest classifier trained on the full SWMS training data reaches accuracy rates of 78.9% on the SWMS test data and of only 74.7% on the SWRK test data. However, when building the classifier directly on the SWRK training data, the same accuracy of 74.7% is reached on the SWRK test set. Thus, the RandomForest classifiers built from the male SWMS training data are just as good in classifying the SWRK test data as their SWRK counterparts. This demonstrates that the lower accuracy in applying the SWMS pitch accent classifier is not due to unsatisfactory generalizability but rather is inherent to the SWRK data. To the contrary, the SWMS classifiers perform just as well as the SWRK classifiers.

A possible explanation for the lower accuracy rates observed for the SWRK data is that the F0 extraction was less reliable for the female data, and that the contribution of the PaIntE attributes derived from the F0 is stronger in predicting pitch accents than in predicting phrase boundaries. This is supported by the observation that if one had to pick just one attribute to decide on the classificiation, the most successful single attribute in classifying boundaries would be the z-score of the final segment, whereas the most successful single attribute in classifying pitch accents would be the PaIntE b parameter.

3.2. Comparison with other studies

Among the recent studies, [10] obtain word accuracy rates of 86.0% for predicting two classes of pitch accents, and of 93.1% for predicting two classes of boundary tones, obtained on data from the Boston Radio News Corpus. Similar results are reported by [11], who obtain slightly lower word-based accuracy rates of 84.2% and 93.0% for two-class pitch accent and boundary prediction on similar data. The accuracies reported by [10] and those by [11] are lower than the best rates achieved in this study, which for the two-class problem were obtained by the LMT alogrithm (86.24% and 93.37%, respectively), although the differences are probably not significant in the boundary case. Comparing our results to the results of these two studies is relatively unproblematic because the data are similar: the two classes of pitch accents and boundary tones are derived from ToBI labels, and the corpora both consist of news-style read speech by professional speakers. However, the corpora are from two different languages with different ToBI systems.

Turning to German data, [12] report accuracy rates of 77.0% for (two-class) pitch accent detection and of 88.6% for (two-class) boundary detection. This is significantly lower than the rates obtained here, but they classify spontaneous user interactions with a wizard-of-Oz system and thus their data is different from the data used in the present study.

Another study reporting German results [13], which also predicts the full set of (English or German) ToBI labels instead of just two classes, achieves syllable-based accuracy rates of 65% and 60% for German and American English pitch accent classification, respectively, and syllable-based accuracy rates of 71% and 68% for German and American English boundary tone classification. We have only provided word-based accuracies in table 1 above, because for the syllable-based data, we have not systematically evaluated all algorithms in several runs. However, taking the RandomForest algorithm, for instance, 10-fold cross-validation in one run yields averaged accuracy rates of 87.5% and 93.9% on the syllable level for pitch accent and boundary tone classification, respectively, predicting the full set in both cases. Thus, our present approach achieves accuracy rates that are considerably higher than those reported in [13].

4. Discussion and Outlook

We have presented results on simulating prosodic categorization using machine learning methods to classify new exemplars. The results compare very well to results reported in other recent studies, particularly to results on German. In contrast to most other studies, our classifiers predict the full set of GToBI labels rather than just two classes.

Various learning schemes implemented in WEKA [1] proved to be equally suitable for prosodic classification showing that good performance in classification is not necessarily due to one outstanding learning algorithm that is particularly suitable for the data; instead, we interpret this as showing that the information provided was sufficient to reliably reach quite high accuracy rates using various learning algorithms.

We have also begun to assess the generalizability of classifiers trained on data of one speaker to other speakers' data. Results showed that the classifiers generalize very well to similar data of another speaker in that they yield the same accuracy rates as classifiers trained directly on data of that speaker.

To pursue the question of generalizing classifiers to data of other speakers, we will apply the best classifiers to data of more speakers, and also to data which do not match the training data so closely with respect to speech style or content.

We also hope to gain some insight into the perceptual relevance of the attributes used in categorizing prosodic events from such simulations. Thus, future work will also involve assessing the contribution that each attribute makes in classifying prosodic events. Attributes that contribute more in that including them serves to significantly raise classification accuracy are hypothesized to be more relevant in the perception of prosody.

5. References

[1] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, USA: Morgan Kaufman, 2005.

[2] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 1998, pp. 311–316.

[3] J. Mayer, "Transcription of German intonation—the Stuttgart system," Institute of Natural Language Processing, University of Stuttgart, Stuttgart, Tech. Rep., 1995.

[4] W. Wahlster, "Smartweb: Mobile applications of the semantic web," in *KI 2004: Advances in Artificial Intelligence*, S. Biundo, T. Frühwirth, and G. Palm, Eds. Berlin/Heidelberg: Springer, 2004, pp. 50 – 51.

[5] Centre for Speech Technology Research, University of Edinburgh, "The Festival text-to-speech synthesis system," "http://www.cstr.ed.ac.uk/projects/festival/".

[6] A. Schweitzer and B. Möbius, "On the structure of internal prosodic models," in *Proceedings of the 15th International Congress of Phonetic Sciences (Barcelona)*, 2003, pp. 1301–1304.

[7] Institute of Natural Language Processing, "IMS German Festival," "www.ims.uni-stuttgart.de/phonetik/synthesis/".

[8] H. Schmid, "Improvements in part-of-speech tagging with an application to German," in *From text to tags—Issues in multilingual language analysis. Proceedings of the EACL SIGDAT Workshop (University College, Belfield, Dublin, Ireland)*, 1995, pp. 47–50.

[9] S. P. Abney, "Chunks and dependencies: bringing processing evidence to bear on syntax," in *Computational Linguistics and the Foundations of Linguistic Theory*. Stanford: CSLI, 1995.

[10] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, 2008.

[11] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, and T. Yoon, "Simultaneous recognition of words and prosody in the boston university radio speech corpus," *Speech Communication*, vol. 46, no. 3-4, pp. 418–439, 2005.

[12] V. Zeißler, J. Adelhardt, A. Batliner, C. Frank, E. Nöth, R. P. Shi, and H. Niemann, *The Prosody Module*. Berlin: Springer, 2006, pp. 139–152.

[13] N. Braunschweiler, "The Prosodizer – automatic prosodic annotations of speech synthesis databases," in *Proceedings of Speech Prosody 2006 (Dresden)*, 2006.