

A Neural Network Model and a Hybrid Approach for Accent Label Prediction

Achim F. Müller^(1,2) and Rüdiger Hoffmann⁽²⁾

(1) Siemens Corporate Technology, Otto-Hahn-Ring 6, D-81739 Munich, Germany

(2) Dresden University of Technology, D-01062 Dresden, Germany

achim.mueller@mchp.siemens.de, kom@eakss1.et.tu-dresden.de

Abstract

In this paper two approaches for data driven prediction of accent labels—perceptual accents and pitch accents—on word level for speech synthesis are presented. In the first approach a causal and retro-causal NN model is used to determine Bayesian *a posteriori* probabilities for the occurrence of a certain accent label. These probabilities are calculated using context windows of part-of-speech (POS) tags and context windows of phrase break labels. In the second approach the probabilities determined by the NN are used as emission probabilities for the states of a Markov model (hybrid approach). The transition probabilities of the Markov model are determined by an *n*-gram. The two approaches are trained and tested on three different prosodically labeled data bases. With both approaches prediction accuracy was higher than that reported in other studies. For qualitative evaluation a new evaluation scheme is presented and discussed. It is found that the first approach applying the NN model gives the best results with respect to the quality of prosodically labeled sentences.

1. Introduction

In many TTS systems the complex task of prosody generation is split into two parts [1][2]. In the first part symbolic prosody labels are generated. These labels are used as input to the second part that generates f0-contours. Symbolic prosody labels can be separated into two types of labels: phrase break labels and accent labels.

In this paper the prediction of accent labels will be discussed. Here the term accent label refers to labels on word level that describe either perceptual prosodic accents or pitch accents that can be derived from the f0-contour. This distinction is made, since the two data bases used in this study apply two different labeling schemes (cf. section 6).

For fast and easy adaptation to new languages and/or speakers a data driven approach is favorable. For symbolic accent label prediction prominent data driven approaches are based on classification and regression trees (CARTs) [3][4]. In [5] and [6] feed-forward NNs and tree-based learning methods have been used to predict word and syllable prominence.

In this paper we present two approaches for accent label prediction. In the first approach a causal and retro-causal NN model is built. In this approach the decision for a certain accent label is based on a context window of surrounding part-of-speech (POS) sequences and phrase break labels. The second approach incorporates the NN model into a Markov model leading to a hybrid approach. This approach enables a global view for a whole sentence. Our system works on sentence level and the goal is to generate a neutral prosody for each sentence as if the sentence was read isolated from surrounding sentences.

The paper is organized as follows: In section 2 the input parameters used for prediction are presented. In section 3 the problem addressed is described by formulars. In section 4 and section 5 the NN model and the hybrid approach, respectively, are presented. We then describe the data bases used for training and testing (section 6). In section 7 results are presented together with a new qualitative evaluation scheme. This scheme is discussed in section 8.

2. Input parameters

Which features are relevant for symbolic accent label prediction had been investigated in [3], [4] and [7]. Generally the following features are used for accent label prediction both on word level and on syllable level [3][6][8]:

- *positional features* that describe at which position the current word or syllable is within an intermediate or intonational phrase (positional features may also describe when the last accent occurred).
- *function/content word features* that describe the distinction between function and content words.
- *POS features* that describe part-of-speech (POS) information.
- *phrase break features* that describe the phrase break information for the current word/syllable and surrounding words/syllables.
- *discourse features* that describe the discourse structure have been proposed in [4].

For the approaches presented in this paper a simple feature set is used that can be easily transferred for the application within a new language. This is important for the use within our multilingual TTS system. Therefore no features for content/function word distinction and no features that describe discourse information are included in the feature set. Further, no positional features are generated.

Thus, the only features used are based on POS (the POS tags used include different tags for punctuation marks, e.g. different tags are used for commas and periods) and phrase break labels for the entire sentence. As phrase break labels, here, only two labels are used denoting whether there is a phrase break after a word or not (no distinction between major and minor breaks). As proposed in [3] and [9], in our system phrase break labels are also predicted for an entire sentence using the method presented in [10] prior to predicting accent labels.

3. Problem Description

The problem of accent label prediction with the input information used (cf. section 2) for an entire sentence can be formulated

as follows:

$$\hat{\mathbf{a}}_m = \arg \max_{\mathbf{a}_m} P(\mathbf{a}_m | \mathbf{c}_m, \mathbf{b}_m) \quad (1)$$

with

$$\mathbf{a}_m = (a_1, \dots, a_m)$$

denoting a sequence of m accent labels for an entire sentence with m words. a_i ($i = 1, \dots, m$) denotes a stochastic variable that describes the accent status of a word. Here, we only distinguish between accented and de-accented words (cf. section 6). Therefore we chose $a_i \in \{0, 1\}$, with $a_i = 1$ meaning word i is accented and $a_i = 0$ meaning the word is de-accented. The variables

$$\mathbf{c}_m = (c_1, \dots, c_m) \quad \text{and} \quad \mathbf{b}_m = (b_1, \dots, b_m)$$

denote sequences of POS tags and sequences of phrase break labels, respectively.

$$\hat{\mathbf{a}}_m = (\hat{a}_1, \dots, \hat{a}_m)$$

denotes the sequence of accent labels for an entire sentence for which the probability is maximized. This sequence needs to be determined.

In the following two sections, two approaches are described, that lead to a sequence $(\hat{a}_1, \dots, \hat{a}_m)$. In the first approach a causal and retro-causal NN is used to determine an accent label for each word by using a window of POS tags and phrase break labels as input. This approach is explained in more detail in [11] and [12]. In the the second approach (hybrid approach) a Markov model is built that enables a global view over the whole sentence.

4. Neural Network Model

Figure 1 shows the block diagram of the causal and retro-causal TDNN used [11][12][13]. As can be seen the system is built up by connecting subnetworks NN_i , $i \in \{-l, \dots, r\}$, $l, r > 0$. The structure of a subnetwork is explained in detail in [11] and [12] and contains a feed-forward NN with shared weights and gating clusters. Each subnetwork is associated with one time step. One time step represents one word within a sentence. The connection of the subnetworks is realized by shared weight matrices \mathbf{A} and \mathbf{A}' (shared weights means to use the same set of weights in each connector denoted by the same upper case bold face letter). \mathbf{A} propagates the causal information flow (from past to future) and \mathbf{A}' propagates the retro-causal information flow (from future to past). The use of shared weight matrices realizes a delay in time which enables the modeling of dynamic processes (also known as finite unfolding in time [14]). Within the architecture of figure 1, shared weight matrices allow a symmetric handling of past and future information. In our application past refers to left context and future refers to right context.

As can be seen in figure 1, each subnetwork¹ NN_i has a vector \mathbf{x}_i as an input and a vector \mathbf{y}_i as output. $\mathbf{x}_i \in \mathcal{R}^m$, $\forall i$ and $\mathbf{y}_i \in \mathcal{R}^n$, $\forall i$, i.e. the input and output vectors, respectively, are of the same dimension for all subnetworks NN_i . The input to the entire system is given by a sequence of input vectors $\mathbf{x}_{t_l}, \dots, \mathbf{x}_{t_0}, \dots, \mathbf{x}_{t_r}$ and the output is given by a sequence of output vectors $\mathbf{y}_{t_l}, \dots, \mathbf{y}_{t_0}, \dots, \mathbf{y}_{t_r}$.

¹Remark on notation of NN in this paper: lower case bold face letters denote signal vectors and upper case bold face letters denote shared weight matrices

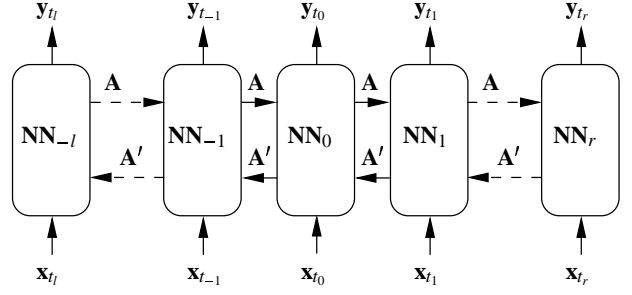


Figure 1: Block diagram of a causal and retro-causal TDNN.

In our application the input vectors $\mathbf{x}_{t_l}, \dots, \mathbf{x}_{t_0}, \dots, \mathbf{x}_{t_r}$ of figure 1 represent a sequence of feature vectors. Feature vectors are determined on word level, i.e. each time step t_i in figure 1 is associated with one word. A feature vector contains the information mentioned in section 1, i.e. POS information and phrase break information for the word at position t_i . An input vector \mathbf{x}_i , $i = -l, \dots, r$, is given by eq. (2):

$$\mathbf{x}_i = (\mathbf{c}'_i, \mathbf{b}'_i) \quad (2)$$

with \mathbf{c}'_i and \mathbf{b}'_i denoting the coded POS tag and coded phrase break label, respectively, for word i . As code a 1-out-of- M code is used. Output vectors, representing the target during training, are also determined on word level. An output vector contains the coded (binary) accent label for each word, as given by eq. (3).

$$\mathbf{y}_i = a'_i \quad (3)$$

By applying a sequence of output vectors as target during training, the NN parameters can be adapted according to the context of surrounding accent labels.

In [15] it is shown that NNs estimate Bayesian *a posteriori* probabilities for class membership, if the following conditions are met: First, the squared-error cost function is used during training and, second, a binary coding is used at the output of the NN. In our case these conditions are met. Therefore, the NN computes the following probability:

$$P_{nn}(a_i) = P(a_i | \mathbf{c}_{lr}, \mathbf{b}_{lr}) \quad (4)$$

with

$$\begin{aligned} \mathbf{c}_{lr} &= (c_{i-l}, \dots, c_i, \dots, c_{i+r}) \quad \text{and} \\ \mathbf{b}_{lr} &= (b_{i-l}, \dots, b_i, \dots, b_{i+r}) \end{aligned}$$

denoting sequences of POS tags and phrase break labels, respectively, with a length of l POS tags (phrase break labels) to the left of i and r POS tags (phrase break labels) to the right of i .

To compute a sequence of accent labels $(\hat{a}_1, \dots, \hat{a}_m)$ for an entire sentence, the windows of length l to the left and length r to the right are shifted across the sentence. Eq. (5) gives one accent label of the sequence.

$$\hat{a}_i = \arg \max_{a_i} P_{nn}(a_i) \quad (5)$$

To get the sequence for an entire sentence, we let $i = 1, \dots, m$.

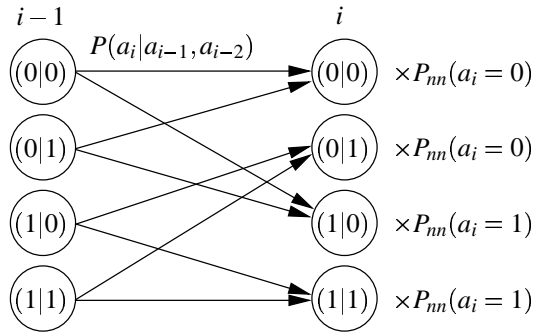


Figure 2: Diagram of a Markov model with 4 states, i.e. $n = 3$ for the underlying n -gram.

5. Hybrid Approach

In the following a Markov model is described where the transition probabilities between states are determined by an n -gram and the emission probabilities are determined by the NN described in the previous section. A Markov model where the emission probabilities are determined by an NN is known as a hybrid approach in speech recognition. The approach is similar to the approach described in [16] for phrase break prediction. The difference lies in the calculation of the emission probabilities. As mentioned, here, the emission probabilities are determined by the NN described in section 4.

The n -gram is used to model the probability of the occurrence of a certain accent label a_i at position i given a sequence of previous accent labels of length $n - 1$:

$$P(a_i|a_{i-1}, \dots, a_{i-n+1}) \quad (6)$$

These probabilities are calculated by counting the number of times each unique sequence of length n occurs in the training data. The probabilities give the transition probabilities for the Markov model. The Markov model has 2^{n-1} states.

With the Markov model, the sequence $\hat{\mathbf{a}}_m = (\hat{a}_1, \dots, \hat{a}_m)$ is given by the following equation:

$$\begin{aligned} \hat{\mathbf{a}}_m &= \arg \max_{\mathbf{a}_m} \prod_{i=1}^m P_{nn}(a_i) \\ &\quad \times P(a_i|a_{i-1}, \dots, a_{i-n+1}) \quad (7) \\ &= \arg \max_{\mathbf{a}_m} \prod_{i=1}^m P(a_i|\mathbf{c}_{lr}, \mathbf{b}_{lr}) \\ &\quad \times P(a_i|a_{i-1}, \dots, a_{i-n+1}) \quad (8) \end{aligned}$$

Figure 2 shows a trellis of a Markov model with 4 states. In this case a 3-gram is used. As indicated, in each state the probability $P(a_i|a_{i-1}, a_{i-2})$ is multiplied with the probability $P_{nn}(a_i)$ determined with the NN. In the forward path only the best path in each state is retained (Viterbi criterion). For each sentence, the backward path with the highest probability is chosen. This way the best possible sequence is determined.

6. Corpora and labeling scheme

For the evaluation of the proposed method, two different corpora with different labeling schemes were used.

The first corpus contains 1000 sentences (21549 words) taken from the German newspaper *Frankfurter Allgemeine Zeitung* [17]. The average sentence length is 21.8 words. The

corpus was read aloud in studio environment by two professional male radio news speakers yielding two spoken versions of the corpus (further referenced by data base FA_{sp1} and data base FA_{sp2}). The corpus is tagged with 35 different part-of-speech (POS) tags. FA_{sp1} was labeled with phrase break labels and accent labels by professional labelers. As phrase break labels three different labels were used, denoting major phrase breaks, minor phrase breaks, and no phrase breaks after each word. The accent labeling scheme used is the same as applied in [18]. Thus the corpus was originally labeled with the following accent labels: PA (perceived primary accent), SA (perceived secondary accent), and EA (perceived emphatic accent). In labeling experiments a low inter-annotator agreement of 70.5% was measured using the above labels. We therefore grouped all accent labels together to one label (PA) and thus only distinguish between words perceived accented and words perceived de-accented. This agrees with experiences made in [3], where it was found that data labeled with two prominence levels was not useful because of the low inter-annotator agreement. With the new labeling scheme an inter-annotator agreement of 86% was achieved for FA_{sp1} . This agreement is still rather low compared to the agreement achieved in other studies (see below). This might be due to the relative monotonous speaking style of the speaker. FA_{sp2} was labeled with minor and major phrase break labels and PA and SA accent labels. As before, the two accent labels were grouped.

The second corpus used is the portion of the *Boston University Radio News* corpus read by the female speaker F2B [19] (further referenced by data base BU_{f2b}). The corpus is labeled with minor and major phrase breaks. For the studies here, words are assigned accent labels describing whether the respective word bears a pitch accent or not. For this labeling scheme an inter-annotator agreement of 91% was reported in [3].

The inter-annotator agreement can to some extent be seen as an upper bound on performance of prediction algorithms.

7. Experiments and Results

For training and testing both corpora have been separated into three subsets which contain approximately the following percentage of data: training set: 70%, validation set: 10% and test set (independent testing data): 20%. All results reported are determined on the test set. The validation set is used to avoid over-fitting to the training data.

Currently our target speaker is represented in the data base FA_{sp1} . Therefore, all of the experiments for qualitative evaluation are done for this data base (see below). The data base FA_{sp2} is used to see if there are significant differences between different speakers. The data base BU_{f2b} is only used for reference.

In our approach the issue of accent placement within the word, e.g. early accent and double accent, is not modeled, since in our German corpus such phenomena occur only for 6% of the words. For the English corpus such phenomena might need special attention [3][9]. However, in this work the focus lies on accent label prediction on word level, which has shown to be a robust basis for prosody generation in our system [13] and for prosody generation across languages in the system used in [6] (even though in [6] prominence is predicted on a scale from 0 to 9, the synthesizer used only distinguishes between accented words and de-accented words). Thus, for the experiments and results discussed below the prediction has been done on word level.

7.1. Results for the Neural Network Model

In the first series of experiments the objective is to see how well the NN performs without the Markov model, i.e. eq. (5) is used to determine a sequence of accent labels $\hat{\mathbf{a}}_m = (\hat{a}_1, \dots, \hat{a}_m)$.

Table 1 displays the results for accent prediction for the three data bases FA_{sp1}, FA_{sp2} and BU_{f2b}. As performance measures three values are displayed. The first value is the overall percentage of correct predictions, i.e. correctly accented and correctly de-accented words (*over*, tab. 1). The second value is the percentage of words predicted to be accented, that should be de-accented, i.e. inserted accents (*ins*, tab. 1). The third value is the percentage of accents that were predicted correctly (*corr(a)*, tab. 1). The three measures have also been used in [16] for evaluation of their method for phrase break prediction. The first three columns of table 1 present results where the originally (hand) labeled phrase break labels were used as input (orig. labeled breaks, tab. 1). The next three columns present results where predicted phrase break labels were used as input (predicted breaks, tab. 1). As mentioned, in our system phrase break labels are generated prior to accent labels for an entire sentence using the method presented in [10]. Phrase break labels are thus available for a whole sentence.

	orig. labeled breaks			predicted breaks		
	<i>over</i>	<i>ins</i>	<i>corr(a)</i>	<i>over</i>	<i>ins</i>	<i>corr(a)</i>
FA _{sp1}	83.1	9.4	81.5	82.6	10.1	82.2
FA _{sp2}	85.8	8.3	85.1	86.1	8.2	85.4
BU _{f2b}	84.5	10.2	90.0	82.9	10.6	88.0

Table 1: Results for accent label prediction for the NN model with different input information for the three data bases.

The overall impression of prediction accuracy is very good. For all three data bases the results for the overall correct rate is higher than the results reported in [3]. In the latter study the data base BU_{f2b} was used and the overall correct rate is reported at 82.5% for the word level. For the studies in [3] the originally labeled phrase break labels were used as an input and not predicted phrase break labels. Thus the results in column one, two, and three of table 1 must be compared to the results of [3]. For a comparison, one needs to consider that the objective in [3] was to predict pitch accents on syllable level. However, the results can to some extent be compared and it is felt that the NN model gives very good results, especially with respect to the achieved quality of the results (see below).

It is surprising that results for the German corpus are in some cases better if predicted phrase break labels are used as an input instead of the hand labeled phrase break labels. The reason for this could be, that the consistency for the predicted phrase breaks is higher and that the NN model can adapt to typical errors in phrase break prediction.

As can be seen from table 1 the rate for insertion errors is quite high, even though the percentage of correctly predicted accent labels marks state of the art accuracy. This holds also true for deleted accents (rates not displayed in tab. 1), i.e. words marked de-accented that should be accented. An important question is whether the insertions and/or deletions occur at positions within a sentence where they disarrange the prosodic structure of the sentence. To answer this question a new qualitative evaluation method is proposed that will be described in the following.

For quality analysis, a prosodically labeled version of the sentences in the test set of the German corpus was generated,

whereby predicted phrase break labels have been used as an input (predicted breaks in table 1) and the NN model had been trained on data base FA_{sp1}. Thus, the resulting testing part of the corpus (177 sentences) includes predicted phrase break labels and predicted accent labels. For this prosodically labeled version of the sentences the overall correct score for phrase break labels is 90.3% (cf. [10]) and the overall correct score for accent labels is 82.6% (cf. tab. 1). The sentences were given to a subject who was asked to rate the quality of the prosodic labeling for each sentence. The subject had previously labeled the data base FA_{sp1} prosodically based on perception which makes him sensitive and critic to possible errors. The subject rated whole sentences as either good, acceptable or bad. The following number of sentences were rated as indicated: 42 (24%) good, 93 (53%) acceptable, and 42 (24%) bad. The applicability of this rating scheme is discussed in section 8.

While rating the sentences the subject marked accents, that were felt to disarrange a prosodic structure that would exist without the marked accent (marked insertion errors). Further, the subject marked words where an accent would be desirable to allow for a better prosodic structure (marked deletion errors). Note that a marked insertion/deletion error does not necessarily have to coincide with an insertion/deletion error observed when comparing the predicted labels with the originally labeled labels.

In general the subject criticized most, that inserted accents disarranged the prosodic structure of the sentences. It was observed that in most cases the marked inserted accents were associated with a tight decision, i.e. the probability given by eq. (4) was close to 0.5. This led to a more conservative accent assignment strategy. A prosodic version of the test sentences was generated where only accents were marked with a probability greater 0.65. This version was again given to the subject. This time the rating was as follows: 70 (40%) good, 83 (47%) acceptable, and 24 (14%) bad. These results for the rating are significantly better compared to the previous version even though the overall correct rate is only 80.3% and thus lower than before.

7.2. Results for the Hybrid Approach

Table 2 shows the results for the hybrid approach using the same measures for performance as for the NN model.

For the experiments related to the results of table 2 two parameters were optimized for each data base. The first parameter is n from the n -gram that determines the probabilities given by eq. (6). For most of the experiments a 3-gram was found to give the best results. The second parameter was introduced to weigh the influence of the NN model used to determine the emission probabilities in the states of the Markov model. Thus in eq. (7) $P_{nm}(a_i)$ is replaced by $P'_{nm}(a_i)$ with

$$P'_{nm}(a_i) = w \times P_{nm}(a_i) \quad (9)$$

The factor w controls the influence of the NN model. w was chosen between 2 and 4 for the experiments of table 2.

As can be seen from table 2 the hybrid approach gives better results in some cases. For data base BU_{f2b} with predicted phrase break labels used as input the highest improvement of 0.6% was achieved. In some cases results were worse than for the NN model.

If the results of table 1 and table 2 are compared, it can be noted that performance for both models ranges at similar high levels. However, both models produce different prosodic structures when applied. The difference of versions lies between 5%

	orig. labeled breaks			predicted breaks		
	<i>over</i>	<i>ins</i>	<i>corr(a)</i>	<i>over</i>	<i>ins</i>	<i>corr(a)</i>
FA _{sp1}	82.0	8.1	75.9	81.5	8.5	75.6
FA _{sp2}	86.3	7.5	84.3	85.9	7.4	83.0
BU _{f2b}	84.2	10.2	89.5	83.5	9.7	87.3

Table 2: Results for accent label prediction for the Markov model with different input information for the three data bases.

and 6%. An important question for us was, whether the quality of the predicted accent labels using the hybrid approach is better than the quality when using the NN model. Therefore, the same qualitative evaluation was performed as for the NN model. Again, the data base FA_{sp1} with predicted phrase breaks was used. The rating was as follows: 29 (16%) good, 75 (42%) acceptable, and 73 (41%) bad. It can be noted that the quality was rated significantly worse than for the NN model.

8. Discussion

It is felt that both data driven approaches presented in this paper to predict accent labels—perceptual accents and pitch accents—on word level give very good results both, quantitatively and qualitatively.

The quantitative evaluation (table 1 and table 2) shows that prediction accuracy on word level is higher for both approaches when compared with the prediction accuracy for the word level reported in [3].

The qualitative evaluation proposed in this paper shows significant differences for different prosodically labeled versions of the test sentences that were generated with the two approaches. The results from the qualitative evaluations are summarized and discussed in the following. A qualitative evaluation was only performed for the results of data base FA_{sp1} with predicted phrase break labels and predicted accent labels.

To get a single measure for each evaluation, we assigned the following values to each sentence with the rating indicated: +1 (the sentence was rated *good*), 0 (the sentence was rated *acceptable*), and -1 (the sentence was rated *bad*). This way mean values for each rating can be calculated, that are presented in table 3. In table 3 NN_{0.5} and NN_{0.65} means that accents are assigned to a word if the probability given by eq. (4) is greater than the indicated value (cf. section 7.1) As can be seen the best

NN _{0.5}	NN _{0.65}	hybrid
0.00	0.26	-0.25

Table 3: Mean ratings from a single subject for the test sentences of data base FA_{sp1}.

results were obtained, if the NN model was used with a conservative strategy for accent assignment (NN_{0.65}).

A very important question when applying the proposed evaluation scheme is whether different subjects rate a given prosodically labeled sentence the same or at least similar. To test this, we let three additional subjects rate the prosodically labeled version of the test sentences that was rated best by the first subject (NN_{0.65} in table 3). From all four subjects (including the first subject), two had previously labeled the speech data bases FA_{sp1} and FA_{sp2} prosodically based on perception. The other two subjects had listened to some of the speech recordings with the appropriately labeled text version. For each of the

four subjects the mean rating is given in table 4. The overall mean for the four subjects is 0.19. This means that the majority of sentences were either found to have a good or acceptable prosodic structure, whereby both, accents and phrase breaks, were predicted.

subj. 1	subj. 2	subj. 3	subj. 4
0.26	0.33	0.11	0.08

Table 4: Mean ratings from four subjects for the test sentences of data base FA_{sp1} for the case NN_{0.65}.

To test whether individual sentences are rated similar by all four subjects, the standard deviation for each sentence rating can be calculated. The mean of the standard deviation for all sentences can be seen as a measure of agreement for the different subjects. The mean of the standard deviation was 0.55. This means that in the mean one to two subjects disagreed in the rating of a sentence with a difference in rating of 1. A difference in rating of 1 means that for example one subject rated *good* and another subject rated *acceptable* (the rating *bad* by the latter subject would mean a difference of 2).

In summary we conclude from the results presented for the qualitative evaluation scheme, that the scheme gives a valid objective measure for the overall quality of prosodically labeled sentences. Since the number of test sentences is very large (177 test sentences), we feel that the good rating for the quality is highly significant.

9. Acknowledgments

We would like to thank Anton Batliner for his support during the labeling of the German corpus.

10. References

- [1] Ralf Haury and Martin Holzapfel, “Optimization of a neural network for speaker and task dependent f0-generation,” in *ICASSP*, 1998.
- [2] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe, “Recent improvements on microsoft’s trainable text-to-speech system — whistler,” in *ICASSP*, 1997.
- [3] K. Ross and M. Ostendorf, “Prediction of abstract prosodic labels for speech synthesis,” *Computer Speech and Language*, vol. 10, pp. 155–185, 1996.
- [4] Julia Hirschberg, “Pitch accent in context: Predicting prominence from text.,” *Artificial Intelligence*, vol. 63, pp. 305–340, 1993.
- [5] Christina Widera, Thomas Portele, and Maria Wolters, “Prediction of word prominence,” in *Eurospeech*, 1997.
- [6] J. W. A. Fackrell, H. Vereecken, J.-P. Martens, and B. Van Coile, “Multilingual prosody modelling using cascades of regression trees and neural networks,” in *Eurospeech 1999*, 1999.
- [7] K. Ross, M. Ostendorf, and S. Shattuck-Hufnagel, “Factors affecting pitch accent placement,” in *ICSLP*, 1992, pp. 365–368.
- [8] Kurt E. Dusterhoff, Alan W. Black, and Paul Taylor, “Using decision trees within the tilt intonation model to predict f0 contours,” in *Eurospeech*, 1999.

- [9] Cameron S. Fordyce and Mari Ostendorf, "Prosody prediction for speech synthesis using transformational rule-based learning," in *ICSLP*, 1998, pp. 843–846.
- [10] Achim F. Müller, Hans Georg Zimmermann, and Ralph Neuneier, "Robust generation of symbolic prosody by a neural classifier based on autoassociators," in *ICASSP*, 2000.
- [11] Achim F. Müller and Rüdiger Hoffmann, "Accent label prediction by time delay neural networks using gating clusters," in *Eurospeech*, Aalborg, Denmark, 2001.
- [12] Achim F. Müller and Hans Georg Zimmermann, "Symbolic prosody modeling by causal retro-causal nns with variable context length," in *Int. Conf. on Artificial NN*, Lecture Notes in Computer Science. Springer, 2001.
- [13] Hans Georg Zimmermann, Achim F. Müller, Çağlayan Erdem, and Rüdiger Hoffmann, "Prosody generation by causal retro-causal error correction neural networks," in *Workshop on Multi-Lingual Speech Communication*, Advanced Telecommunications Research Institute International (ATR), 2000.
- [14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart and J. L. McClelland, Eds., vol. I, pp. 318–362. MIT Press/Bradford Books, Cambridge, MA, 1986.
- [15] Michael D. Richard and Richard P. Lippmann, "Neural network classifiers estimate bayesian a posteriori probabilities," *Neural Computation*, vol. 3, no. 4, pp. 461–483, 1991.
- [16] Alan W. Black and Paul Taylor, "Assigning phrase breaks from part-of-speech sequences," in *Eurospeech*, 1997.
- [17] Institut für Phonetik und sprachliche Kommunikation, "Siemens synthese korpus - si1000p," corpus available at <http://www.phonetik.uni-muenchen.de/Bas/>.
- [18] A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann, "Automatic annotation and classification of phrase accents in spontaneous speech," in *Eurospeech*, 1999.
- [19] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Tech. Rep., ECS Department, Boston University, Boston, MA, 1995.