Optimization of MFNs for signal-based phrase break prediction

Michael Hofmann and Oliver Jokisch

Laboratory of Acoustics and Speech Communication
Dresden University of Technology, 01062 Dresden, Germany
{michael.hofmann3,oliver.jokisch}@mailbox.tu-dresden.de

Abstract

The automatic prosodic annotation of large speech corpora gains increasing consideration since appropriate databases for the training of prosodic models in speech synthesis and recognition are needed. On linguistic level, correct phrase and accent marking are essential processing steps. The authors developed a neural network based method for signal-based phrase break prediction and tested this method across two different speech databases.

The structure of the multilayer feed-forward neural network (MFN) had been optimized and adapted to the target database and to the specific annotation task. The method is rather data sensitive—depending on different human labelers and small differences across training databases, like frequency of occurrence or strength of phrase breaks. The MFN method can be easily adapted to the characteristics of different databases (long or short phrases, special formats like dates or web addresses, etc.). If applied to different databases which contain phrase markers of human experts, phrase break recognition rates vary from 79 % up to 97 %.

1. Introduction

Automated speech processing requires large speech databases to enable corresponding variability of natural speech, to provide appropriate training material for algorithms and, in general, to enhance speech quality. Besides large acoustic inventories for corpus-based speech synthesis, there is also an increasing demand for prosodically annotated speech corpora in research and for commercial use since the correct modeling of prosodic features has a strong effect on the achievable overall speech quality in speech synthesizers.

With respect to necessary databases in several languages, appropriate automatic tools for segmentation and annotation are required. The prediction of speech parameters can be implemented by rule-based or data-driven approaches. A MFN is directly trained on the available speech data and can therefore be easily adapted to different training databases and application scenarios.

In the study, the neural network input vector is derived from signal based and linguistic features that can be obtained by standard signal processing and by available preprocessing modules of speech synthesis. Neural networks can deal with the resulting extensive feature vectors, focusing on the required and ignoring the redundant features [3]. The study describes the optimization of MFNs for the phrase break prediction. A similar MFN optimization was already used in the prediction of suitable Fujisaki parameters for the f0 contour modeling [10].

Neural networks pose difficulties to introspection since the weight matrix of the network hides the modeling process. Nevertheless, it is possible to analyze and to optimize external aspects like the training process, the network structure or the input feature vector.

2. Database and annotation

The TC-Star project aims at producing speech corpora that can be used for building advanced state-of-the-art TTS systems as well as for intralingual and interlingual research on voice conversion and expressive speech [7]. It is going to provide high-quality language resources for UK English, Spanish and Mandarin. The voices recordings are sampled at 96 kHz and with 24 bit precision and each voice is subdivided into several corpus parts: Novels and short stories are included as well as expressive speech.

The required volume of 10 h of net speech per voice corresponds to about 90 000 running words that need to be extensively annotated. Because of the sheer volume of data most of this task has to be performed automatically.

The data analyzed in this report are built from single sentences and short paragraphs read by the UK voice *rob*. The training data consist of 158 sentences with 2 344 words containing 3 596 syllables. This corpus solely consists of well-defined read speech and does not contain spontaneous utterances.

The necessary prosodic transcription includes the detection of phrase breaks and pitch accents. Both features have been annotated using two levels: phrase breaks are divided into *minor* (intermediate intonational phrases) and *major* breaks (full intonational phrases). The pitch accents can be *normal* or *emphatic*.

An example sentence and its base frequency contour is shown in figure 1. It includes the following required prosodic markup for the description of pitch accent and phrase structure:

> As regards <b> *nitrogen*# levels, <BB> we would need reliable *statistics*# <b> and *data*# from the various Member *States*#. <BB>

A word followed by # is marked as normally, the one followed by ## as emphatically accented. Intermediate intonational phrases are delimited by <b>, whereas major phrase breaks are marked with <BB>. The used corpus contains only well-defined read speech and is therefore only labeled with one level of stress.

3. Empirical neural network topology

So far, neural network theory does not provide determinate rule to obtain optimal network structure, number of hidden layers, transfer functions, etc. for a particular task (compare also [11]). The employed neural network structure in this study is derived from similar experiments in pattern recognition and prediction and led to good results for a variety of problems.

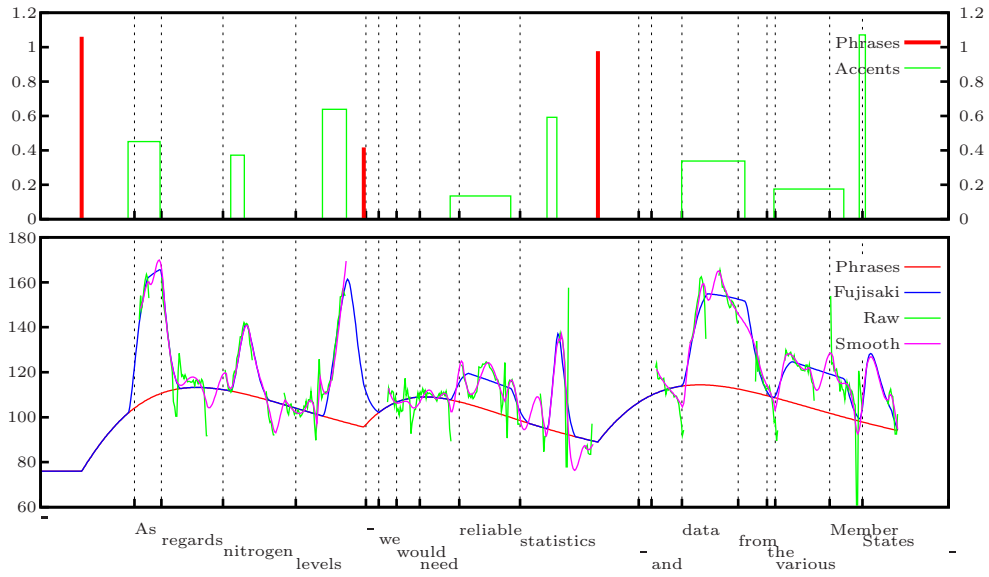

Figure 1: Example sentence of the TC-STAR corpus *rob200*

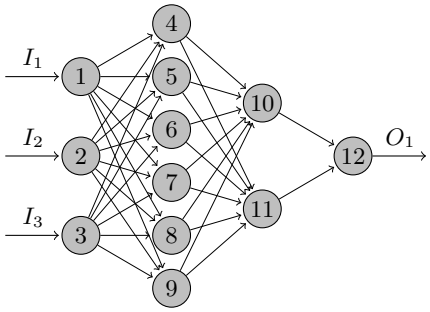

Figure 2: Multilayer feed-forward neural network with three input neurons, six and two neurons in the two hidden layers and one output neuron

The prediction is done by a multilayer feed-forward network with a variable number of neurons in the two hidden layers (figure 2). The input unit number is equal to the size of the feature set, and one output neuron for the requested parameter is used. All neurons are realized by a sigmoid activation function and are trained with backpropagation.

The output value for each pattern is set to a value between zero and one, depending on the number of classes for the training. The training pattern set is grouped by category and patterns are duplicated as necessary to get groups of equal strength.

4. Feature vectors

A large number of input features for the prediction of prosodic parameters have been proposed which model fundamental frequency, duration, energy and linguistic information on syllabic and word level [1], especially

- F0 onset, offset and linear regression coefficients
- Absolute and normalized duration
- Energy contour linear regression coefficients
- Part of speech (POS) tags [8]
- Rule-based boundary labels. [5]

Additionally to these features, the implemented algorithm is based on the following assumptions:

1. Only word-level features are used, as syllable-level features do not provide additional performance according to [2].
2. A configurable context window provides the network with information of previous and following words. The best window size has to be determined by optimization.
3. If possible, both normalized and absolute values are provided. Although normalized variables seem to serve intuitively better as network input, they may be seriously outperformed by absolute values in some cases [4].

The following input fields are provided:

Duration Several different features describing the duration of the current word are provided:

- Absolute word duration.
- Normalized word duration as calculated from the mean phoneme durations of the whole corpus.
- Scale factor comparing absolute and normalized word durations.

Linguistic information The output of the Festival speech synthesis system [6] is leveraged to provide information about POS tags and possible phrase structure:

- The POS calculated by the Festival HMM tagger is considering the *Penn Treebank* notation with 45 classes.
- A derived POS tag which uses only 16 simplified classes.
- Phrase break information from the probabilistic Festival tagger that is enabled for American and British English voices included in Festival. It uses the probabilities of breaks based on the POS of the previous and following words of a break combined with an ngram model of the break distribution to optimize the phrasing.

Base frequency and power contour The logarithmic f0 and power contours $A(t)$ are described with the following features [8], they are calculated for raw as well as for smoothed contours where applicable.

- First signal value A_{on}. For f0 contours, the first value of the first voiced segment is used.
- Last signal value A_{off}. For f0 contours, the last value of the last voiced segment is used.
- Maximum and minimum value A_{min} and A_{max}.
- Mean value A_{mean}, for f0 contours only voiced segments are included.
- Absolute and relative position of the maximum and minimum value t_{min} and t_{max}.
- Linear regression coefficients m and n and residual sum of squares R for the contour. For f0 contours, only voiced segments are considered.

Fujisaki parameters The well-known Fujisaki model (introduction e. g. in [9]) provides parameterization of f0 contours according to physical and physiological aspects. It is available for many languages, e. g. Korean, Mandarin or German [13]. The f0 contour is modeled by Fujisaki phrase and accent commands which should correspond to perceived prosodic features. To extract word level features related to the Fujisaki parameterization, the smoothed f0 contour is processed by an automatic Fujisaki parameter extraction tool [12]. Afterwards, the following input components are calculated:

- Accumulated accent area (score) per word for all Fujisaki accent commands.
- Phrase command amplitude if there is a phrase command in the current word or in a possibly following silence.
- Maximum accent score per Fujisaki phrase as derived from the Fujisaki phrase commands.

Structural information Certain additional positional and structural components are provided. Phrase segmentation for these features is obtained from the phrase break annotation of the Festival synthesis.

- Number of syllables in the current word, current phrase and previous phrase.
- Index of the current word in the phrase and utterance.
- Index of the first syllable of the current word in the phrase and utterance.
- Start and end time of the current word and phrase.

5. Network training and results

Different network structures and training scenarios have been evaluated to determine the performance of the described approach. The analysis concentrates on the following questions:

- How well can phrase breaks be recognized? Does the result improve if the phrase break categories for intermediate and final breaks are combined into one class?
- Is the approach suitable for predictive use, i. e. can the labeling be inferred by a network that is only trained on a small part of the whole corpus?
- Is the resulting network independent from specific databases or speakers?

C	N_1	N_2	RR_{NB}	RR_{B}	RR_{BB}	RR	$\overline{RR}$
0	15	10	94.8	52.6	71.1	88.9	72.8
1	15	10	97.3	40.4	86.7	90.9	74.8
2	10	6	97.5	45.6	95.6	92.2	79.6
2	15	10	98.1	40.4	93.3	92.1	77.3

Table 1: Recognition rates for 75 % training set and separate break classes (C: context size, N_1, N_2: number of neurons in the hidden layers)

C	N_1	N_2	RR_{NB}	$RR_{\text{B\|BB}}$	RR	$\overline{RR}$
0	15	10	96.0	80.6	93.3	88.3
1	15	10	96.3	86.4	94.6	91.4
2	10	6	96.0	77.7	92.8	86.9
2	15	10	97.5	75.7	93.7	86.6

Table 2: Recognition rates for 75 % training set and merged intermediate and major break classes (C: context size, N_1, N_2: number of neurons in the hidden layers)

5.1. Phrase break recognition

The resulting recognition rates for the *rob200* corpus of a 75 % training set extracted from all patterns and with separated as well as with merged intermediate and major break categories are displayed in table 1 and 2. The columns RR_{NB}, $RR_{\text{B|BB}}$, RR_{B} and RR_{BB} contain the recognition rates for words followed by no break, any break, an intermediate break or a major phrase boundary, respectively. RR denotes the overall recognition rate independent from the break category and $\overline{RR}$ the arithmetic mean of all recognition rates for the single classes.

Two network structures with different number of neurons N_1 and N_2 in the two hidden layers and various context sizes C are compared. Because of the varying recognition results of different classes for one network, $\overline{RR}$ is always lower than RR. Improved context knowledge as well as an increased number of neurons generally leads to better recognition rates. The largest network with ± 2 words context size seems to generalize worse, fitting more of the noise in the data set than the smaller networks. The overall recognition rate for merged categories is about 2 %, the averge class recognition rate about 12 % better than the rate observed for separate classes.

5.2. Phrase break prediction

Table 3 compares the results of a training with a 75 % training set to one with a training set of only 25 % of all patterns, thus modeling a partially hand-labeled corpus where the rest should be annotated automatically. The results are only slightly worse for the 25 % training set and the network seems to be able to derive all the specific prosodic features influencing the phrase segmentation for a given speaker from a small part of the whole corpus (25 % corresponds to about 40 sentences).

5.3. Speaker dependency

To test the speaker dependency of the trained network, a totally different database is used for comparison purpose. Table 4 shows a summary of the results for the *rob* and the *kate* corpora. The *kate* corpus from a female speaker consists of 1 197 short phrases that contain mainly names, dates, numbers and web addresses.

			RR		$\overline{RR}$	
C	N_1	N_2	75 %	25 %	75 %	25 %
0	15	10	93.3	92.6	88.3	89.3
1	15	10	94.6	93.0	91.4	88.0
2	10	6	92.8	92.3	86.9	86.1
2	15	10	93.7	92.9	86.6	85.7

Table 3: Recognition rates with 75 % and 25 % training sets and merged intermediate and major break classes (C: context size, N_1, N_2: number of neurons in the hidden layers)

	RR_S	RR_M	$\overline{RR}_S$	$\overline{RR}_M$
kate 75 %	80.3	78.6	83.8	83.2
kate 25 %	79.2	77.2	80.8	80.2
rob 100 %	71.1	70.0	77.7	75.1

Table 4: Recognition rates of the *kate* corpus for a neural network with 15 and 10 neurons in the hidden layers and a context size of ± 2 words. It includes training sets for split and merged break classes. (Test patterns are completely different from training patterns.)

The average recognition rate for a network which is trained on the *rob* and used for the *kate* corpus is about 3 % to 6 % worse than for the one trained on the *kate* corpus itself.

6. Conclusion

The authors developed a neural network based method for signal-based phrase break prediction and tested this method across two different speech databases. The structure of the MFN was optimized and adapted to the specific target database and to the specific annotation task within TC-STAR project. The automatic annotation results depend on different human reference labelers and small differences across training databases.

Nevertheless, MFN method can be easily adapted to the characteristics of different databases. The recognition performance for a network trained on first database and tested on the second independent database is only about 3 % to 6 % worse than performing the experiment on first database only. In realistic application scenario one can expect phrase break recognition rates from about 79 % up to 97 %.

The authors will further study potential speaker, database but also language dependencies of the proposed MFN method and will extensively compare this method with other existing rule-based and data-driven approaches.

7. References

[1] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nõth, and H. Niemann. Prosodic feature evaluation: brute force or welldesigned? In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey, editors, *Proceedings of the 14th International Congress of Phonetic Sciences*, volume 3, pages 2315–2318, San Francisco, California, USA, 1999.

[2] A. Batliner, J. Buckow, H. Niemann, and V. Warnke. The prosody module. In W. Wahlster, editor, *Verbmobil: foundations of speech-to-speech translation*, pages 106–121. Springer, 2000.

[3] A. Batliner, A. Kießling, R. Kompe, H. Niemann, and E. Nöth. Can we tell apart intonation from prosody (if we look at accents and boundaries)? In G. Kouroupetroglou, editor, *Proceedings of the European Speech Communication Association (ESCA) Workshop on Intonation: Theory, Models and Applications*, pages 39–42, Athens, Greece, 1997.

[4] A. Batliner, E. Nöth, J. Buckow, R. Huber, V. Warnke, and H. Niemann. Duration features in prosodic classification: why normalization comes second, and what they really encode. In M. Bacchiani, J. Hirschberg, D. Litman, and M. Ostendorf, editors, *Proceedings of the International Speech Communication Association (ISCA) Tutorial and Research Workshop Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, pages 23–28, Red Bank, New Jersey, USA, 2001.

[5] A. Batliner, V. Warnke, E. Nöth, J. Buckow, R. Huber, and M. Nutt. How to label accent position in spontaneous speech automatically with the help of syntactic-prosodic boundary labels. Verbmobil-Report 228, 1998.

[6] A. W. Black, P. A. Taylor, and R. Caley. *The Festival speech synthesis system*. University of Edinburgh, 2002.

[7] A. Bonafonte, H. Höge, H. S. Tropf, A. Moreno, H. van der Heuvel, D. Sündermann, U. Ziegenhain, J. Pérez, I. Kiss, and O. Jokisch. TTS baselines and specifications. Technology and Corpora for Speech to Speech Translation (TC-STAR) Deliverable D8, 2004.

[8] J. Buckow, A. Batliner, R. Huber, H. Niemann, E. Nöth, and V. Warnke. Detection of prosodic events using acoustic-prosodic features and part-of-speech tags. In *Proceedings of the 5th International Workshop Speech and Computer (SPECOM)*, pages 63–66, St. Petersburg, Russia, 2000.

[9] H. Fujisaki. The interplay between physiology, physics and phonetics in the production of tonal features of speech of various languages. In G. Kokkinakis, N. Fakotakis, E. Dermatas, and R. Potapova, editors, *Proceedings of the 10th International Workshop Speech and Computer (SPECOM)*, volume 1, pages 39–48, Patras, Greece, 2005.

[10] M. Hofmann. Optimization of a data-driven prosody generation. Research paper, Dresden University of Technology, 2004. (In German).

[11] O. Jokisch and M. Hofmann. Evolutionary optimization of an adaptive prosody model. In Soon Hyob Kim and Dae Hee Youn, editors, *Proceedings of the 8th International Conference on Spoken Language Processing*, pages 797–800, Jeju Island, Korea, 2004.

[12] H. Kruschke and M. Lenz. Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2881–2884, Geneva, Switzerland, 2003.

[13] H. Mixdorff. *Intonation patterns of German - model-based quantitative analysis and synthesis of f0 contours*. PhD thesis, Dresden University of Technology, 1997.