Data driven symbolic prosody modeling

Janez Stergar, Bogomir Horvat, Zdravko Kačič

University of Maribor,

Faculty of Electrical Engineering and Computer Science Maribor, Slovenia

E-mail: janez.stergar@uni-mb.si, bogo.horvat@uni-mb.si, zdravko.kačič@uni-mb.si

ABSTRACT

In this paper a data-driven approach in prediction of word level prosody breaks will be presented. A semi-automatic approach of symbolic prosody raw text annotation will be introduced together with a module for prediction of annotated tags for Slovenian language. The prediction module is based on an automatic learning technique, which depends on the construction of a large corpus labeled appropriately. This labeling can be done either automatically, or by hand. While automatic labeling can be less accurate than hand labeling, the latter is very time consuming and, in many cases, inconsistent. Therefore, we constructed an interactive tool for word level prosody annotation (major/minor breaks) as well as annotation of prominent words. Our major concern was speeding up the process of manual annotation with an increase of consistency and required time effort minimization. In parallel a semi-automatic approach for determining prosody breaks and prominent words for read speech was developed with the goal to support different speakers and new languages in the future. The labeled Slovenian corpus has been used to train our phrase break prediction module, implementing a neural network (NN). Experiments for the data-driven prediction of major=minor and major/minor phrase breaks were performed. The prediction accuracy achieved marks state-of-the-art word level prosody breaks prediction for the Slovenian language and is comparable to the prediction accuracy of other approaches in which more complex NN structures or other prediction methods were applied, and a much larger corpus was used for training.

1. INTRODUCTION

The general problem of speech synthesis is to produce the most intelligible and natural audio stimuli from a set of linguistic, paralinguistic and non-linguistic (e.g. emotional) instructions. The most important parameter of speech, having a great impact on naturalness and intelligibility is prosody [1].

What is prosody? We can define prosody as the systematic organization of various linguistic units into an utterance or a coherent group of utterances in the process of speech production. Its realization involves both segmental and suprasegmental features of speech, and serves to convey not only linguistic, but also paralinguistic and non-linguistic information [2].

As prosody involves sound attributes that encompass larger units such as syllables, words, sentences and even paragraphs, our concern will be focused on two important aspects concerning speech prosody: phrasing and prominence. Phrasing refers to the way in which sentences and text are segmented into smaller entities, while prominence refers to the strength of relation between elements within a given domain [13]. Proper modeling of the two mentioned parameters makes the syntactic and semantic structure of the message more transparent to the listener (prosodic models).

The problem of producing good prosody models can be tackled either by using: linguistic expertise – adapting the models by hand, or automatic learning techniques to adapt the models automatically by making use of large speech corpora.

The second approach offers the potential for rapid model adaptation and can, to some extent, be seen as language independent [5].

Improvement in prosody parameter prediction remains a challenge for producing highly natural text-to-speech systems (TTS). Data-driven techniques allow prosodic regularities to be extracted automatically from a prosodic database of natural speech. Furthermore, if the potential exists for rapid development of prosody models, it then becomes possible to develop several prosody models easily, each suited to a specific type of speech synthesis. Nevertheless, automatic learning techniques offer a solution when adapting prosodic models to a new language, voice or a new application, which can be an advantage in multilingual TTS environments.

In the beginning an inventory for the data-driven approach of symbolic prosody modeling has to be created. First, a suitable corpus has to be designed. Second, towards the creation of appropriate prosodic models, annotation (labeling) of the corpora with appropriate prosodic tags is necessary.

The labeling can be performed by hand, or automatically. While automatic labeling can be less accurate than hand labeling, prosodic labeling based on perceptual tests is time-consuming and usually inconsistent. People with expert phonetics and linguistics knowledge are required.

ISBN 1-876346-48-5

2. DATA DRIVEN TECHNIQUES

As manual labeling is time and cost intensive, automatically labeling approaches in the design of databases are preferred [14; 8]. They also allow rapid adaptation to new languages and/or databases and, therefore, are suitable for multilingual approaches.

The first methodological problem is how to obtain relevant training material as each speech segment is subject to interaction of many components: coartculation effects, perceptual constraints, structure of the language, speaker's commitment, mood and listener's expectations and nevertheless SNR [1]. Must the reference corpus really have a sufficient statistical coverage of the constraints mentioned? Not necessarily, if the reference corpus reflects the intended characteristics of the target speech [1]. The modeling of such characteristics can thus be considered as an information retrieval process – data driven.

The final step in symbolic prosody modeling is the transformation of predicted symbolic tags into acoustic realization. Usually the mapping is realized within the pitch generation algorithm. There are different intonation modeling approaches, which differ in the way describing prosody and can be classified into: single point defined modeling (INTSINT), two points defined modeling (ToBI), multi point modeling (TILT) and fully specified modeling. The last of the four mentioned modeling approaches (Stem-ML) is one of the promising in next generation of TTS systems [6]. Stem-ML provides a well-defined mapping from tags to f0 contours, replacing the pitch generation algorithm of the TTS system.

3. BUILDING A PROSODIC MODEL FOR SLOVENIAN LANGUAGE

Building a prosodic model for Slovenian language

A prototype of an interactive tool for semi-automatic symbolic prosody annotation was developed. This tool (with a GUI) combines the advantage of hand labeling and automatic labeling by achieving a high consistency in labeling and reducing the time needed for hand labeling. In our approach the use of a graphic tool for prosody labeling minimized the required expert knowledge. Our goal was to reduce manpower, time, and expenses for prosodic labeling. The tool has a graphical interface helping the labeler (expert or novice) to consistently label symbolic phrase boundaries and, therefore, minimize the time required for the procedure [12].

The database

For the training material we used a corpus designed for concatenative speech synthesis for Slovenian language [10]. The corpus consists of 1206 preselected sentences (app. three hours of speech). These texts have been designed to ensure good coverage of phones, therefore, no intentional balancing of clause types was performed (declarative – interrogative – exclamations), dialogue context and syntax were not considered, and no semantic analysis was performed since only isolated sentences were included. We are aware that the fact of sparsity in such corpus could be a problem, nevertheless the problem of sparsity in such reference corpus is perhaps ill-posed [1].

The audio database recordings were created in a studio environment with a male speaker reading aloud-isolated sentences in the Slovenian language (sampled at 44.1 kHz /16 bit).

Corpus annotation

In the first step the text corpus was hand-labeled with part-of-speech (POS) tags. We used a very restricted inventory of POS tags minimized to nouns, verbs, adjectives, adverbs, numbers (ordinal + cardinal), pronouns, prepositions, conjunctions, particles, interjections and two different tags for punctuation. All tags were combined in an environment where tracking and correcting tags was simplified for the labelers [12].

In the second step prosody breaks were annotated semi-automatically using our graphic tool. We used tags similar to those used in [7]. The following inventory of break tags was applied:

- B3 – full intonational boundary with strong intonational marking, often with lengthening or change in speech tempo (we'll refer to that tag as a major break);
- B2 – intermediate phrase boundary with weak marking (we'll refer to that tag as a minor break);
- B9 – irregular prosodic boundary, usually disfluencies at hesitations, repairs etc. ; and
- B0 – normal word boundary.

The last, third step, focuses on prominent words annotation. The experiments made were focused only on pitch accent annotation within a phrase using two tags (PA, NA). In our inventory a phrase is a sequence of words within B2/B3 tags. Distinction between accented (PA) and non-accented (NA) words were done comparing syllable pitch envelope and pitch dynamics (syllables mean average pitch changes normalized on syllable mean average changes of pitch for the concerned sentence). Energy and mean energy for syllables in each word were also considered. Word prominence was classified according to four classes similar to those used in [3]:

- EA = Emphatic accent,
- PA = Primary accent,
- SA = Secondary accent, and
- NA = No accent.

ISBN 1-876346-48-5

4. PHRASE BREAK PREDICTION MODULE

Which parameters are relevant for symbolic prosody label prediction remains an open research question. A feature set that is commonly used and seems to be relatively independent of language and task is part-of-speech (POS)

The dimension of our input vector (NN) as well as tag set is similar to the German language prediction tests as reported in [9], ((4+1+4) * tag set = 117).

The task of our NN model was to predict phrase breaks using a commonly used tag set that was relatively independent of language and task. A basic NN model of a Multi Layer Perceptron (MLP) with one hidden layer was used. The output parameters were set to 1 or 0 to mark the presence or absence of a major=minor (major/minor) and non-break phrase boundary. All nodes of the NN used in our experiments have a sigmoid logistic activation function. All input parameters have discrete values.

A variation of the standard back-propagation algorithm the so-called VarioEta [11] was applied to train the NN. Patterns from input vectors were selected with a quasi-stochastic procedure. The initial setting for the training started with a learning rate of η=0.05, decreasing its value every 10 epochs. No extensive weight and node pruning was performed.

5. EXPERIMENTS AND RESULTS

Annotation and prediction experiments were performed for prosody breaks. In the first experiment prosody break labels were marked at positions indicated by the tool. This resulted in database TLDB (Tool Labeled Data-Base). In the second experiment, a careful analysis of the f0 contours, energy contours, and perception led to the insertion of break labels at positions that the tool did not indicate. This annotation scheme resulted in a database referenced as AHLDB (Additionally Hand Labeled DataBase). The increase of B2 tags in AHLDB compared to TLDB is proportional for almost all POS tags. The increase of B9 labels is small relative to the increase of B2 labels and in our opinion, is strongly speaker dependent.

The complete annotation of 600 clauses had been performed and it was possible to detect over 93% of major phrase breaks (over 77% of all breaks) and considerably shorten the time needed for annotating the database when the semi-automatic method was used.

The two databases, AHLDB and TLDB, were used to train the phrase break prediction module of our TTS system. For both databases the B9 labels marking hesitations were removed prior to training, since hesitations occur generally at positions where a break seems unsuitable. Both databases were split identically into a training set (70% of the data), a validation set (10% of the data) used to avoid overfitting, and a generalization set (20% of the data). All results were determined on the independent generalization set.

Tests were made comparing the prediction results using labeled data in AHLDB and TLDB as training data for our phrase break prediction module. The results were significantly better for TLDB (Table 1). This probably is due to the fact that only pauses were taken into account for prediction. It seems that the rhythm (frequency of occurrences) of short and long pauses produced by the speaker (interpreting the different clauses) seems to be consistent and can be predicted with a very high accuracy. Nevertheless, the consistency of automatic annotation also should be emphasized. We also showed that the prediction accuracy decreases only insignificantly for the case in which minor and major breaks were not grouped (minor/major) if TLDB was used to train the phrase break prediction module (Table 1).

Table 1. Comparisons of phrase break prediction for AHLDB and TLDB.

breaks	AHLDB	TLDB
minor/major	74.05 %	79.37 %
minor=major	77.74 %	80.60 %

This means that for the case of break vs. non-break prediction, this tool can be used for annotation without much performance loss. This fact results in a significant reduction in the time needed to label a database.

Considering these results and the fact that tests of prediction accuracy, regardless of the approach used for prediction, are preferably made with hand-labeled corpora [4; 9], we decided to conduct further experiments with the AHLDB.

Table 2. Results for phrase break prediction AHLDB.

breaks	B correct	NB incorrect	Overall
AHLDB	77.74 %	4.95 %	94.03 %

In Table 2 the result is presented for the phrase break prediction module using AHLDB database for training. Despite the limited annotation material available (600 labeled clauses were used) and MLP NN structure used for prediction, the results are comparable to those for German [9] and English [4]. For the prediction of breaks (B correct), the results are equivalent to the achieved accuracy prediction of B correct (77.67 %) for German in [9] and nearly equivalent to the achieved accuracy prediction of B correct (79.27 %) for English in [4], despite a much smaller inventory of clauses used.

Slightly better overall and non-break (NB incorrect) prediction accuracy as reported for German [9], and for English [4], was achieved. The prediction results are promising, but it is evident that there are still problems in the percentage of breaks being predicted as non-breaks (Table 3 non-breaks row, breaks column).

Table 3. Confusion matrix for the generalization data.

predicted/actual	breaks	non-breaks	all predicted
breaks	2008	256	2264
non-breaks	582	11176	11758
all actual	2615	11432	

ISBN 1-876346-48-5

The amount of breaks being falsely predicted as non-breaks is much larger compared to the amount of breaks being falsely predicted as non-breaks (currently ten times more non-breaks are being falsely predicted as breaks compared to breaks being falsely predicted as non-breaks).

6. CONCLUSION

This paper presents an approach for the annotation and modeling of symbolic prosody tags – phrase breaks for the Slovenian language. A tool for annotating the corpora with prosodic tags semi-automatically was designed and used for the labeling of a Slovenian corpus. This tool can be seen as a step towards the semi-automatic annotation of prosody features for the Slovenian language.

The aim was also the design of an environment for annotation and prediction of prosody features with an approach suitable also for other languages. Therefore, we used only those features for prosody prediction that were relatively independent of language and task.

Our conclusion is that the approach we used is very useful for prosody annotation of the Slovenian language. First, it considerably reduces the time needed for annotation and second, it provides a high level of support to the labeler that facilitates consistent annotation of prosodic events.

The database for the Slovenian language labeled using the proposed tool was used to train our phrase break prediction module [12]. The NN structure used for prediction is a good compromise when compared to other, more sophisticated, NN structures and approaches to the prediction of symbolic prosody markers (complexity). On the other hand, it is necessary to balance the relatively high ratio of breaks being predicted as non-breaks to non-breaks being predicted as breaks. This could be managed with the use of more sophisticated prediction structures less susceptible to unbalanced input data flow (over four times more non-breaks than breaks, in the case of AHLDB).

The achieved prediction accuracy marks the state-of-the-art for phrase break prediction accuracy for the Slovenian language. A minor clause inventory was used, compared to other approaches, with equivalent or superior success in phrase break prediction accuracy.

A subjective quality analysis of the prosodic phrases of test sentences was performed. The results from this analysis indicate a high potential for the practical application of the annotation tool in combination with our phrase break prediction module.

REFERENCES

[1] Bailly G. Introduction to Prosody in Speech Synthesis. Y. Sagisaka, N. Campbell, N. Higuchi editors. Computing Prosody. Computational Models for Processing Spontaneous Speech. Springer–Verlag New York. 1997.

[2] Fujisaki H. Prosody, Models, and Spontaneous Speech. Y. Sagisaka, N. Campbell, N. Higuchi editors. Computing Prosody. Computational Models for Processing Spontaneous Speech. Springer–Verlag New York. 1997.

[3] Bavarian Archive of Speech SI1000 (1998). Prosodic Markers Version 1.0 University of Munich, Institute of Phonetics. Munich, Germany.

[4] Black A. W., Taylor P. (1997). Assigning Phrase Breaks from Part-of-speech Sequences. Proceedings Eurospeech 97, pp. 995-998. Rhodes, Greece.

[5] Fackrell J. W. A., Vereecken H., Martens J.-P., Van Coile B. (1999). Multilingual Prosody Model-ling using Cascades of Regression trees and Neuronal Networks. Proceedings Eurospeech 99, vol. 4, pp. 1835-1838. Budapest, Hungary.

[6] Kochansky G., Shih C. Prosody Modeling with Soft Templates. Speech Communication. Elsevier Science. Volume 39. Issues 3-4, p.p. 311-352.

[7] Kompe R. Prosody in Speech Understanding Systems. Springer – Verlag Berlin Heidelberg, Lecture Notes in Artificial Inteligence, 1997.

[8] Malfrere F., Dutoit T. and Mertens P. Fully automatic prosody generator for text-to-speech. ICSLP 98, pp. 1395-1398. Sydney, Australia.

[9] Müller A. F., Zimmermann H.G., and Neuneier R. Robust Generation of Symbolic Prosody by a Neural Classifier Based on Autoassociators. Proceednings ICASSP 00, vol. 3., pp.1285-1288. Istanbul, Turkey.

[10] Rojc M., Kačič Z. Design of Optimal Slovenian Speech Corpus for use in the concatenative Speech Synthesis System, pp. 321-325. LREC 00, Athens, Greece, 2000.

[11] SENN. Version 3.0 User Manual. SIEMENS AG. 1998

[12] Stergar J. Determining Symbolic Prosody Features with analysis of Speech Corpora. Master Thesis. University of Maribor. Faculty for EE. and Comp. Sci. 2000.

[13] Terken J., Collier R.. The Generation of Pro-sodic Structure and Intonation in Speech Synthesis. In Speech Coding and Synthesis, Kleijn W.B. et all Edts. pp. 635-662. Elsevier 1998.

[14] Vereecken98 Vereecken H., Martens J. P., Grover C., Fackrell J., Van Coile B. Automatic prosodic labeling of 6 languages. ICSLP 98, Sydney Australia.