Designing Prosodic Databases for Automatic Modeling of Slovenian Language in a Multilingual TTS System

Achim F. Müller[(1)], Janez Stergar[(2)], Bogomir Horvat[(2)]

(1) Siemens Corporate Technology, Dept. CTIC 5, 81730 Munich, Germany
(2) University of Maribor, Fact. for EE and Comp. Science Maribor, Slovenia
achim.mueller@mchp.siemens.de, {janez.stergar, bogo.horvat}@uni-mb.si

Abstract

In this paper the design of a prosodic data base and the data driven prediction of phrase breaks for modeling Slovenian language in a multilingual text-to-speech (TTS) system are presented. Automatic learning techniques offer a solution in adapting prosodic models to a new language, voice or a new application, because they allow prosodic regularities to be automatically extracted from a prosodic database of natural speech. Such techniques depend on the construction of a large corpus labeled with symbolic prosody labels. The 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. Therefore an interactive tool for semi-automatic labeling that uses the segmented spoken counterpart of the text as input will be presented. The tool combines the advantage of hand labeling and automatic labeling by achieving a high consistency in labeling and reducing the time that would be needed for hand labeling. The labeled Slovenian corpus has been used to train our phrase break prediction module. Experiments for the data driven prediction of major and minor phrase break labels have been performed. The achieved prediction accuracy marks state-of-the art for phrase break prediction accuracy for Slovenian language.

1. Introduction

Improvement in prosody prediction remains a challenge for producing really natural text to speech systems (TTS). As manual labeling is time and cost intensive automatically labeled databases are preferred (Vereecken, 1998; Malfrere, 1998).

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 of rapid model adaptation and can to some extent be seen as language independent (Fackrell, 1999).

Data driven approaches allow rapid adaptation to new languages and/or databases and therefore are suitable for multilingual approaches where large speech corpora are processed and models for prosody generation are adapted.

Prosodic labeling based on perceptual tests is very time consuming and usually inconsistent. Man force with expert phonetics and linguistics knowledge is required. In this paper we will present an approach, implementing Slovenian language in a multilingual TTS system and suggest the use of a tool to minimize the required expert knowledge for prosodic labeling. The goal is to reduce man force, 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 minimizes the time required for labeling.

2. Database

To our knowledge, there exist no prosodically labeled corpora for Slovenian language that can be used for prosody research in speech synthesis. An important step during the adaptation of the system to the Slovenian language was therefore the design of a suitable database.

2.1. The Corpus

The corpus consists of 1206 sentences in Slovenian language (orthography) which equals app. 3 hours of speech. To our knowledge this is the largest prosodically annotated database for Slovenian language. Therefore we would like to emphasize the pioneer work accomplished in building a corpus of national importance.

The selection of the text was designed to ensure good coverage of the phones in the Slovenian language, also some clauses were gathered and included from different text styles (e.g. literature and newspaper text).

The texts were not chosen to achieve good coverage of prosodic patterns, i.e. no balancing of clause types was performed, dialog context and syntax was not considered and no semantic analysis was performed since there are only isolated sentences included – prosody was not relevant for text selection.

The whole corpus was determined with the selection of clauses from 31 million words corpus in Slovenian language from e-newspapers, e-literature on the WWW or CD's. The major part of clauses covers daily-published news and the minority consists of clauses taken from Slovenian classics and poetry.

The majority of sentences in the database have between 15 and 25 words. Four different text corpora were selected and statistically analyzed. The selection of sentences for the final corpus is based on a two-stage process. In the first stage an analysis based on statistical criteria is performed. In the second stage the final text is chosen based on the result of the first stage. In the following the two stages are described

The statistic analysis of corresponding units generated in a non-uniform units generator is performed on the sentence level for each of the four corpora in a separate module. The module sorts all units and determines how frequently each unit appears. The obtained statistics mirrors the non-uniform unit richness of all clauses and unit structure for the corresponding text corpora (Rojc, 2000).

After the described statistical analysis of the text the final corpus was generated. The criterion for final text filtering was based on monophones, diphones, triphones and fivefones richness.

The corpora that were statistically analyzed have similar unit statistics, although distributions of units across the corpora were not the same (Table I). A careful elimination of sentences considering comprehension and frequency of units was performed. Clauses with poor unit comprehension and unit duplicates were reduced. In the final corpus 1200 sentences remained. The corpus is used to adapt our multilingual TTS system to Slovenian language.

Text corpus	1	2	3	4
Monophones	38	38	38	38
Diphones	1030	1001	1028	1028
Triphones	11398	10233	11283	9126
Fivephones	64811	55218	69668	52039

Table I: Statistical analysis of phones

2.2. Audio Recordings

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

Since the speaker is a professional radio news speaker the speech contains no disfluencies (i.e. filled pauses, repetitions and deletitions) although for this particular speaker there are some indices for hesitations in form of pauses and lengthening. Compared to the German corpus of resembling extent used in our TTS system (Institut für Phonetik und sprachliche Kommunikation, 1998) the percentage of hesitations differs essentially.

2.3. Phonetic Transcription

The phonetic transcription was managed with a two step conversion module. The first step is realized with a rule-based algorithm. Subsequent to the first step the second step was designed with a data driven approach (neural networks were used).

The module was designed for the support of two approaches in grapheme–to–phoneme conversion. The first part was intended for the case that no morphological lexica were available. First rule based stress assignment is done, followed by grapheme-to-phoneme conversion procedure.

The step of stress marking before grapheme-to-phoneme conversion is very important for Slovenian language, since it very much depends on the type and place of the stress. If the phonetic lexicon is available, a data driven approach, representing the second part in the module, using neural networks can be used. Here, the phonetic lexicon is used as a data source for training the neural networks.

The data preparation, the generation of the training patterns and the training of neural networks are done completely automatically. The transcription is performed in two steps. In the first step the graphemes are converted into phonemes and syllable breaks are inserted in the phoneme string. In the second step stress marks are inserted. The problem how to perform mapping between

```
Dvesto NUM deset NUM centimetrov SUBST visoki ADJ Nemec SUBST B2 ne ADV skriva VERB ambicij SUBST v PREP ameriški ADJ ligi SUBST B3 , PUNC saj ADV je VERB tik ADJ B2 pred PREP prvenstvom SUBST zavrnil VERB nekaj PRON ponudb SUBST B2 bogatih ADJ evropskih ADJ klubov SUBST B3 . PUNC
```

Figure 1. With POS and prosody breaks labeled clause

graphemes and phonemes by generation of training patterns for neural networks, was solved as proposed in (Hain, 1999). For both NN tasks we applied a multilayer perceptron (MLP) feed-forward network with one hidden layer. As learning algorithm the back-propagation algorithm was chosen (Rojc, 2000).

The pronunciation is derived from the IPA-Alphabet. In order to represent the IPA symbols in ASCII characters the SAMPA format is widely used. In our grapheme-to-phoneme conversion module the SAMPA phonetic transcription symbols for Slovenian language are used (Rojc, 2000).

2.4. Part of Speech Tags

The text corpus was hand labeled using the following simplified part-of-speech tags (POS):

1. SUBST for nouns,
2. VERB for verbs,
3. ADJ for adjectives,
4. ADV for adverbs,
5. NUM for ordinal and cardinal numbers,
6. PRON for pronouns (nouns and adverbs),
7. PRED for predicative,
8. PREP for prepositions,
9. CONJ for conjunctions,
10. PART for particle,
11. INT for interjection and
12. PUNC for punctuation.

All tags are combined in an environment were tracking and correcting tags is simplified for the labelers (Stergar, 2000).

Compared to the tag-set for our German corpus (Institut für Phonetik und sprachliche Kommunikation, 1998), the tag-set for Slovenian language is smaller. The difference in size occurs because the Slovenian corpus was hand-tagged and no reliable tagger currently exists for a large tag-set (Figure 1).

2.5. Phonetic Segmentation and Labeling

The spoken corpus was phonetically transcribed using HTK. Along standard nomenclature we used two special markers for pauses between phonemes. “sil” denotes the

```
#!MLF!#
"*/stavek_1.lab"
0 1750000 sil
3900000 5000000 s
5000000 5250000 t
5250000 5550000 O
5550000 5550000 sp
```

Figure 2. An example of phonetic segmented and labeled text

silence before and after sentence. “sp” denotes the silence between words in the sentence. We determined both with one state HMM and all phonemes with three state HMM in the HTK environment (Figure 2).

3. Prosody Labeling

Since no symbolic prosody labels are defined for Slovenian language, we decided to use similar labels as used in (Kompe, 1997). Thus prosodic labels are determined through acoustic perceptual sessions and text is labeled speaker dependent. The following labels were used for labeling the corpus:

- B3 prosodic clause/phrase boundary
- B2 prosodic phrase boundary
- B9 irregular prosodic boundary, usually hesitation, lengthening and unwanted pauses and
- B0 for every other boundary.

The acoustic prosodic boundaries were determined by boundary indication, listening to audio files and visual output (pitch and energy) from our tool (cf. section 3.1 and section 3.2).

3.1. A Tool for Semiautomatic Prosody Labeling

We designed a tool (Figure 3) intended to help the labeler (novice or expert) to make decisions about prosody breaks within each sentence. The tool indicates possible prosody boundaries, which depend on the segmented pauses in spoken corpora.

Experiments on multilingual databases (3 languages) have shown that the strategy of segmenting the speech signal with pauses yields a significant improvement of the annotation accuracy (Vereecken, 1997).

Syllable and word boundaries are marked by vertical lines adding overview clearness and *B* marks for symbolic prosody boundaries are inserted in the sentence concerned (Figure 3).

The tool indicates marks for prosody boundaries taking phonetic segmentation of pauses into account. The position of prosody boundaries is selected considering time of silence between words. The decision of indication is made on comparison with a specific threshold. This threshold can be changed manually (Stergar, 2000).

3.2. Labeling Results

Two labeling experiments have been performed. In the first experiment, labels were marked at positions indicated by the tool. Additionally, a careful analysis of the f0-contours and energy contours and perception has lead to the insertion at positions that the tool did not indicate. This labeling scheme resulted in a database further referenced by DB1.

In the second experiment, labels were only marked at positions indicated by the tool. This resulted in database DB2.

The frequencies of the occurrence for each labeled break for each POS tag are presented in table III. The increase of B2 tags in DB1 compared to DB2 is proportional for almost all POS tags. The increase of B9

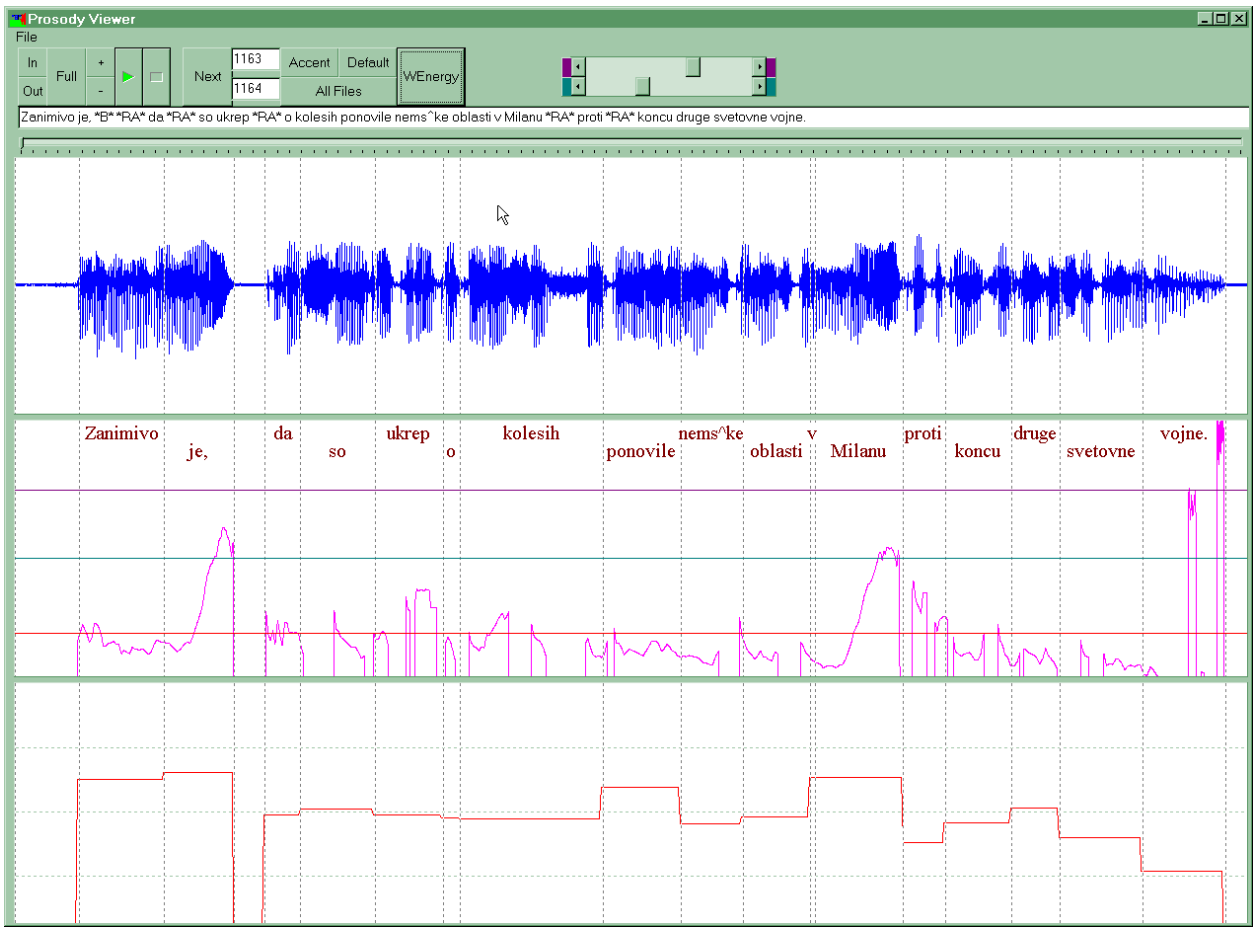

Figure 3. The indication of symbolic prosody boundaries

POS	B3_DB2	B3_DB1	B2_DB2	B2_DB1	B9_DB2	B9_DB1	Δ (%)
SUBST	595	627	886	1266	101	101	42,89
VERB	97	103	172	260	109	128	51,16
ADJ	25	26	65	99	24	29	52,3
ADV	28	28	50	91	63	70	82
PRON	10	10	22	35	36	51	59,09
NUM	5	6	14	25	13	6	78,57
CONJ	0	0	5	5	46	60	0
INT	0	0	5	5	0	0	0
PREP	0	0	0	0	27	46	0
PART	0	1	0	0	0	0	0

Table III: The increase of B2, B3 and B9 tags (Δ) concerning POS tags in DB1 and DB2

labels is evidently minor to the increase of B2 labels and in our opinion is strongly speaker dependent.

Until now we performed the complete labeling for half the corpus. With the semiautomatic method used we were able to detect 77,95% of all breaks (over 93 % for B3) (Stergar, 2000) and considerable shorten the time needed for labeling the database (over 50%).

4. Prosody Prediction

After labeling the corpus as described in the preceding sections, the two databases (DB1 and DB2) were used to train the phrase break prediction module (Müller, 2000) of our text-to-speech system. For both databases the B9-labels marking hesitations were removed prior to training, since the hesitations generally occur at positions were a pause seems not suitable.

Both databases were identically split 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 reported below are determined on the independent generalization set. The reported figures represent overall accuracy, i.e. correctly classified non-breaks are included in the statistic.

breaks	DB1	DB2	Comparison
minor/major	91.34%	94.27%	89.87%
minor=major	91.75%	95.23%	91.67%

Table II: Results for phrase break prediction

Table II presents the results for prediction of minor and major breaks (line 1) and for prediction of only major breaks (line 2). For the latter experiments, minor breaks were grouped with major breaks. If no breaks are predicted at all, prediction accuracy is 81.6%. This can be seen as a baseline for worst case prediction.

As can be seen from table 2, the results are significantly better for DB2. This is probably due to the fact that the consistency for the labels detected by the tool is much higher than the consistency for the case where additional breaks are labeled based on perception.

In the third column the predicted phrase breaks after training on DB2 are compared to the phrase breaks originally labeled in DB1. As can be seen prediction accuracy degrades only insignificantly for the case were minor and major breaks were grouped (minor = major) if DB2 was used to train the phrase break prediction module. This means that for the case of break vs. non-break prediction the tool can be used for labeling without performance loss. This results in a significant reduction in time needed to label a database.

For quality evaluation, the sentences from the test set (121 sentences) with automatically generated phrase breaks have been subjectively rated as good, acceptable and bad. The prosodic phrasing was found to be good for 95 sentences (78.5%), acceptable for 23 sentences (19%) and bad for only 3 sentences (2.5%).

5. Conclusion

We presented a tool for labeling and classification of prosody breaks. This tool can be seen as a first step towards automatic labeling of prosody breaks of Slovenian language. We conclude that our approach in using the segmented pauses in the speech corpus for prosody boundary indication for Slovenian language is very useful. Firstly, it considerably reduces the time needed for labeling and, secondly, it provides a high level of support to a labeler for consistent labeling of prosodic phrase boundaries.

The data base for Slovenian language labeled with the proposed tool was used to train our phrase break prediction module (Stergar, 2000). The achieved prediction accuracy marks state-of-the art for phrase break prediction accuracy for Slovenian language. 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 labeling tool in combination with our phrase break prediction module.

6. References

Fackrell J. W. A., Vereecken H., Martens J.-P., Van Coile B., 1999. Multilingual Prosody Modelling using Cascades of Regression trees and Neuronal Networks. Proceedings EUROSPEECH 99, Budapest, Hungary.

Hain H.-U., 1999. Automation of the training procedure for neural networks performing multilingual grapheme to phoneme conversion, Proceedings EUROSPEECH 1999, Budapest, Hungary.

Institut für Phonetik und sprachliche Kommunikation. (1998) Siemens Synthese Korpus - SI1000P, http://www.phonetik.uni-muenchen.de/Bas/.

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

Malfrere F., Dutoit T. and Mertens P., 1998. Fully automatic prosody generator for text-to-speech. ICSLP 98, Sydney Australia.

Müller A. F., Zimmermann H. G., Neuneier R., 2000. Robust Generation of Symbolic Prosody by a Neural Classifier Based on Autoassociators, ICASSP 2000.

Rojc M., Kačič Z., 2000. Design of Optimal Slovenian Speech Corpus for use in the concatenative Speech Syn-thesis System. LREC2000 Athens Greece.

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

Stergar J., Hozjan V., 2000. Steps towards preparation of text corpora for data driven symbolic prosody labelling. T. Erjavec, J. Gros, (edt.). Language technologies: proceedings of the conference. Ljubljana, 2000.

Vereecken H., Martens J. P., Grover C., Fackrell J., Van Coile B., 1998. Automatic prosodic labeling of 6 languages. ICSLP 98, Sydney Australia.

Vereecken H., Vorstermans A., Martens J. –P. and Van coile B., 1997. Improving the Phonetic Annotation by means of Prosodic Phrasing. EUROSPEECH 97, Greece.