Steps towards preparation of text corpora for data driven symbolic prosody labelling

Janez Stergar, Vladimir Hozjan

Faculty of Electrical Engineering and Computer Science
Institute of Electronics
University of Maribor
Smetanova 17, SI-2000 Maribor, Slovenia
{janez.stergar, vladimir.hozjan}@uni-mb.si

Abstract

Automatic learning techniques offer a solution in adapting prosodic models to a new voice or a new application, because they allow prosodic regularities to be automatically extracted from a prosodic database of natural speech. Such techniques unfortunately depend on the construction of a large corpus, which is generally hand-labeled. This labeling process is extremely time-consuming, inconsistent (labeler depended) and is an obstacle in the process of rapidly adapting the prosody. Usually for language specific labeling a special language knowledge is needed. A step towards preparing labeled speech corpora with prosodic labels was taken. A simple tool, constructed for interactive annotation of prosody events using the segmented spoken counterpart of the raw ASCII text file, as input, will be presented. An annotator accepts or rejects the proposed prosody break. Yet no classification methods for boundary strength have been implemented.

1. Introduction

Improvement in prosody prediction remains a challenge for producing really natural text to speech systems (TTS). As manual labeling suffers from some major drawbacks, therefore usage of automatically labelled databases is suggested in literature (Vereecken et al., 1998, Malfrere et al., 1998).

Two approaches in determining the symbolic prosodic markers are common (Fackrell, 1999);

- linguistic expertise - hand crafting the models for large speech corpora and
- automatic learning techniques.

The second approach has attracted much interest, since it offers the potential of rapid model development and language independence.

The potential of rapid model development and language independence lies in data driven approach where large speech corpora are processed and models for prosody generation are derived.

Prosodic labeling based on perceptual test is very time consuming and usually inconsistent. Man force with expert phonetics and linguistics knowledge is required.

In our approach we suggest the use of a tool to minimize the required expert knowledge with the goal of reducing man force, time and expenses in designing text corpora for data driven approaches in symbolic prosody labeling.

Therefore we decided to build a simple symbolic prosody boundary detector helping the labeler (expert or novice) for consistency in symbolic boundary labeling and minimizing time constraints.

2. Prosody and TTS

Prosody may be defined as the ensemble of sound attributes that do not constitute the phonetic identity of individual speech segments, but rather encompass larger units such as syllables, words, sentences and even paragraphs. Prosody, therefore, concerns the suprasegmental aspects of spoken language: variations in pitch, duration, loudness and timbre.

Whereas phrasing and prominence seem to be universal properties of language, their phonological manifestation is strongly language-specific. This makes it necessary to study for each individual language the phonological forms that are associated with these phenomena. Furthermore, the functions associated with phrasing and word prominence may differ across languages, and this will affect the kind of rules to be devised for the assignment of prosodic structure for individual language (Kleijn et al., 1998).

2.1. Generation of Prosodic structure

Generally in TTS systems prosody is generated in two main steps. First an abstract symbolic description is derived on the basis of syntactic structure and second the symbol string is converted into an acoustic description of prosody (phoneme duration, f0 curve and energy contour).

2.2. Symbolic prosody boundaries

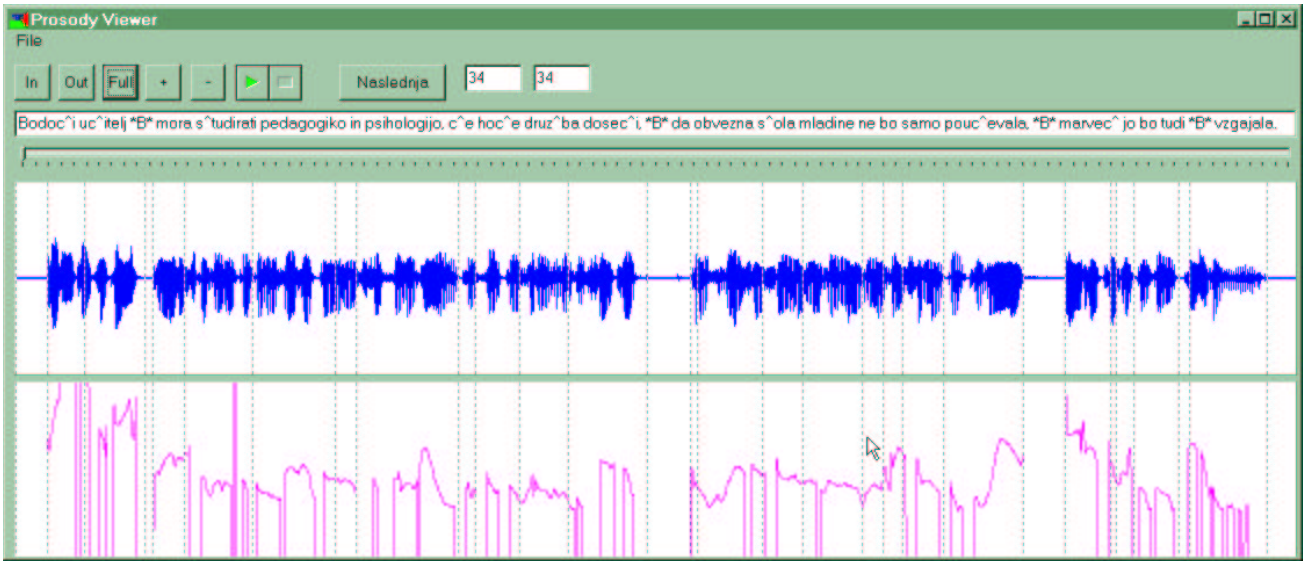
Generally two symbolic prosodic classifications are used;

- syntactic prosodic boundaries and
- acoustic prosodic boundaries.

In the first class boundaries are concerned which are determined from rules or recommendations. Unfortunately usually only recommendations are available.

Although it is generally accepted that there is no one-to-one mapping between syntactic and prosodic constituency, it is also evident that syntactic information is a necessary prerequisite to the issue of prosodic phrasing. Some of the syntactic boundaries coincide with prosodic boundaries, but not all do. In addition the hierarchical structure is completely different (Klejin 1998).

```
             this is the cat that ate the rat that sat on the mat
syntactic:  (        (       (        (       (        (         ))))))
prosodic:   (               )(               )(                  )
```



Bodoc^i uc^itelj *B* mora s^tudirati pedagogiko in psihologijo, c^e hoc^e druz^ba dosec^i, *B* da obvezna s^ola mladine ne bo samo pouc^evala, *B* marvec^ jo bo tudi *B* vzgajala.

Figure 1: The indication of symbolic prosody boundaries.

2.3. Used prosody markers

The labels usually used are;

- M3 for clause boundaries,
- M2 for constituent boundaries likely to be marked prosodically,
- M1 that are close to a M3 boundary and are most certainly not marked prosodically and
- M0 for every other word boundary.

The second class is determined through acoustic perceptual sessions and text is labeled speaker dependent. Acoustic-prosodic boundaries were defined as in Institute of Phonetics (BAS) and used in VERBMOBIL (Kompe, 1997);

- B3 prosodic clause 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 listening, visual output and boundary indication from our tool. Also the passages on indicated boundaries were taken into the consideration.

3. Database construction

3.1. Text Corpus

The text corpus consists of 1206 text sentences in Slovenian language (orthography) which equals app. 3 hours of speech. The selection of the text was designed to ensure good coverage of the phones in the language also some clauses gathered from different text styles were included (literature). Other textual factors that might be prosodically important were not considered.

The whole corpus was determined with gathering 31 million words from e-newspapers, e-literature on the WEB or CD's.

Sentences were chosen between 15 and 25 words. 4 different text corpora were generated and statistically analyzed. After statistically analyzing the text the final corpora was generated. The criteria for final text filtering was based on monophones, diphones, triphones and fivefones richness.

Table 1 : Statistical analysis of phones.

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

With the goal not to loose any unit from the corpus only wanting to remove the units duplicates, a careful elimination of sentences considering the unit context was performed. At the end all sentences with duplicated units were removed (Rojc, 2000).

3.2. Audio Recordings

The audio database recordings were created with a male speaker reading aloud isolated sentences sampled at 44.1 KHz (16bit).

3.3. Phonetic transcription

The orthographic transcription was managed with a two step conversion module. The first step is rule-based and the second step is data driven (neural networks were used). The first part was intended for the case were no morphological lexica was 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 the latter very much depends on the type and place of the stress. In case we have available phonetic lexicon, data driven approach, representing the second part in the module, using neural network can be used. Here, the phonetic lexicon is used as a data source for training the neural networks. The neural network which was taken for the basis of this part is based on a method used and described in SNNS.

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. The first one converts the graphemes into phonemes and inserts the syllable breaks in the phoneme string, and the second one inserts stress marks. The problem how to perform mapping between graphemes and phonemes by generation of training patterns for neural network, was solved automatically.

For both neural networks a multilayer perceptron (MLP) feed-forward network with one hidden layer was used. As learning algorithm the back-propagation algorithm was chosen.

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).

3.4. Part of Speech labels

The text corpus was hand labeled using the following simplified Part of Speech markers (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 (noun and adverb)
7. PRED for predicative
8. PREP for prepositions
9. CONJ for conjunctions
10. PART for particle
11. INT for interjection
12. PUNC for punctuation

An example of POS labeled sentence:

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

3.5. Phonetic segmentation and labeling

The spoken corpus was phonetically transcribed using HTK.

Phone durations

```
#!MLF!#
"*/stavek_1.lab"
0 1750000 sil
1750000 2650000 d
2650000 2950000 v
2950000 3900000 e:
3900000 5000000 s
5000000 5250000 t
5250000 5550000 O
5550000 5550000 sp
5550000 6000000 d
6000000 6600000 E
```

First line denotes that this database is an output of HTK. In second line is the name of the file from which was calculated begin and end of phonemes. All other lines point to the beginning and ending of phonemes. First column denotes phoneme beginning, the second phoneme ending in the speech signal file. The third column denotes the name of the phoneme. There are two special markers for pauses between phonemes. “sil” denotes the silence before and after sentence. “sp” denotes the silence between words in the sentence.

3.6. Boundary strength labeling

Yet we used only three labels; B2, B3, B9 as described in preceding paragraphs.

4. A tool for interactive prosody indication

The tool (Figure 1) was intended to make indications of possible prosody boundaries depending on the detected pauses in spoken corpora.

Syllable and word boundaries are line marked adding overview clearness and *B* marks for symbolic prosody boundaries are inserted in the sentence concerned.

Tool inserts marks for symbolic prosody boundaries concerning phonetic segmentation. The position of symbolic prosody boundary is selected where is the silence between words is larger then before selected threshold. This threshold can be changed manually.

Yet we did not implement any boundary strength classification and syntactical prosody labels were not discussed.

4.1. Prosody boundaries labeled text

An example of hand labeled text:

Bojevnika sta po rivalstvu pisno naznanila premirje B3, potem ko sta med B9 rokovanjem v Jerebovi jedilnici pozirala B2 ob ogromnem polnjenem B2 morskem psu B3.

The symbolic labels for acoustic boundaries described in preceding section were used.

An example of semiautomatic labeled text corpora for the same sentence:

Bojevnika sta po rivalstvu pisno naznanila premirje, *B*
potem ko sta med *B* rokovanjem v Jerebovi jedilnici
pozirala *B* ob ogromnem polnjenem *B* morskem psu.

Yet only breaks with no classification are indicated.

5. Results

The comparison between the semiautomatic and hand-labeled text corpora was done.

Table 2: Comparison between semiautomatic and hand-labeled corpora.

	Number	%
B correct	2779	89,72
B incorrect	0	0
B2 incorrect	260	8,39
B3 incorrect	9	0,29
B9 incorrect	49	1,58
Overall B incorrect	318	10,27
Overall B	3096	

The results are promising in comparison due tests done with other only hand-labeled corpora in Data-Driven Generation of Symbolic Prosody (Müller, 2000).

6. Future work

The phonetic transcription should be additionally handcrafted-checked by an expert minimizing the errors in symbolic boundary detector.

Extending the proposed tool with classification algorithms and adding syntactic prosodic boundaries in classification scheme will be the main topic of our future work. Furthermore adding word prominence labels and finally automatic generation of boundary strength is also our future goal.

The completing of corpora should be done adding word prominence labeled corpus and some tests of importance of non linguistic expertise in POS and prosodic boundary strength labeling should be done. We would like to determine the importance of wrong/right labeling the different POS classes in addition to automatic learning approaches.

7. References

C. Grover, J. Fackrell, H. Vereecken, J.-P. Martens, B Van Coile (1998). *Designing Prosodic Databases for Automatic Modelling in 6 languages*. Third ESCA/COSCODA Workshop on Speech Synthesis, Australia.

Matej Rojc, Zdravko Kačič (2000). *Design of Optimal Slovenian Speech Corpus for use in the concatenative Speech Synthesis System.* LREC2000 Athens Greece.

R. Kompe, (1997). *Prosody in Speech Understanding Systems*. Springer – Verlag Berlin Heidelberg, Lecture Notes in Artificial Inteligence 1307.

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

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

W. B. Kleijn, K. K. Paliwal. *Speech Coding and Synthesis. Elsevier*. 1998.

J. W. A. Fackrell, H Vereecken, J.-P. Martens, B. Van Coile. *Multilingual Prosody Modelling using Cascades of Regression trees and Neuronal Networks*. 6th European Conference On Speech Communication And Technology September 5-9, 1999 Budapest, Hungary.

A. F. Müller, Janez Stergar. *Data-Driven Generation of Symbolic Prosody for Slovene Language*. Advances in Speech Technology. International Workshop Maribor 2000.