PREDICTION OF SYMBOLIC PROSODY BREAKS WITH NEURAL NETS

Janez Stergar, Bogomir Horvat

University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia

Key words: symbolic prosody, data driven approach, prosodic breaks, symbolic prosodic tags, prediction, Neural Networks, Multi Layer Perceptrons, DSP

Abstract: In this paper the data driven prediction of word level prosody modeling for Slovenian language is presented. Automatic learning techniques depend on the construction of a large corpus labeled with appropriate 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 and in some cases inconsistent. Therefore we will present a new semi-automatic approach for determining prosody breaks features with a graphical user interface (GUI). The GUI 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, using a neural network (NN) structure. We used an MLP structure suitable for Digital Signal Processor (DSP) implementation. Experiments for the data driven prediction of major/minor phrase breaks have been performed. The achieved prediction accuracy marks a good entry-level for phrase break prediction of the Slovenian language and is comparable to other approaches in phrase break prediction where more complex prediction methods were used and a much larger corpus was used for training. The achieved overall prediction accuracy is about 90 %.

Napovedovanje simboličnih prozodičnih mej z nevronskimi mrežami

Ključne besede: simbolična prozodija, podatkovno vodeni pristop, prozodične meje, simbolične prozodične značnice, napovedovanje, digitalni signalni procesor, nevronske mreže, večplastni perceptroni

Izvleček: V članku bomo predstavili podatkovno vodeno modeliranje prozodije na nivoju besed za slovenski jezik. Samodejne tehnike učenja so odvisne od zasnove obsežne besedilne zbirke označene z ustreznimi značnicami. Označevanje lahko izvedemo samodejno ali ročno. Čeprav je samodejno označevanje ponavadi manj natančno kot ročno, predstavlja slednje časovno zelo obsežno proceduro, ki je v določenih primerih nedosledna. Predstavili bomo postopek polavtomatskega določanje prozodičnih mej z uporabo interaktivnega grafičnega vmesnika (GUV). GUV združuje prednosti ročnega s samodejnim, bolj konsistentnim označevanjem in prispeva k zmanjšanju potrebnega časa za označevanje. Označena besedilna zbirka v slovenščini je bila uporabljena pri učenju modula za napovedovanje prozodičnih mej. Modul smo zasnovali na strukturi nevronskh mrež z večplastnimi perceptroni, ki je primernejša za implementacijo v digitalnih signalnih procesorjih. Izvedli smo poskuse za napovedovanje večjih/manjših prozodičnih mej. Dosežena uspešnost napovedovanja predstavlja dobro izgodišče pri napovedovanju prozodije na nivoju besed za slovenski jezik in je primerljiva z drugimi pristopi napovedovanja prozodičnih mej, kjer so za napovedovanje uporabljene bolj kompleksne metode in strukture ter bistveno obsežnejše besedilne zbirke za učenje. Spošna uspešnost napovedovanja mej (je/ni) presega 90%.

1 Introduction

Automatic learning techniques offer a solution when adapting prosodic models to a new language (in a multilingual text-to-speech (TTS) system), voice or a new application. Data driven techniques allow prosodic regularities to be automatically extracted from a prosodic database of natural speech. Such techniques depend on the construction of a large corpus labelled with symbolic prosody labels.

In the first steps toward creating an inventory for the data driven approach of symbolic prosody prediction, the labelling of data can be performed by hand as well as automatically if no reference corpora is available. While automatic labelling can be less accurate than hand labelling, the latter is very time consuming.

Both goals had to be accomplished: the preparation of hand labelled corpora and in parallel the development of automatic labelling techniques to somehow speed up the labelling process. Therefore a prototype of a new, interactive GUI (tool) for semi-automatic symbolic prosody labelling, which uses the segmented spoken counterpart of the text as input was developed. This tool combines the advantage of hand labelling and automatic labelling by achieving a high consistency in labelling and reduces the time needed for hand labelling.

Improvement in prosody prediction remains a challenge for producing really natural text to speech systems (TTS). As manual labelling is time and cost intensive, automatically labelled databases are preferred /6/, /17/.

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 /2/.

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 are adapted for prosody generation.

Prosodic labelling based on perceptual tests is very time consuming and usually inconsistent. People with expert phonetics and linguistic knowledge are required. In the presented approach, avoiding the necessary expert, the use of a graphic tool to minimise the required expert knowledge was proposed. Our goal was to reduce manpower, time, and expenses for prosodic labelling. The tool has a graphical interface helping the labeller (expert or novice) to consistently label symbolic phrase boundaries and, therefore, minimise the time required for labelling.

2 Database

To our knowledge, no prosodically labelled corpora for Slovenian language exists, that can be used for prosody research in speech synthesis. An important step during the adaptation of a TTS system to a new language is the design of a suitable database.

2.1 The corpus

The corpus consists of 1206 sentences in the 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 Slovenian language, also some clauses were gathered and included from different text styles (e.g. literature and newspaper texts).

The majority of sentences in the database had between 15 and 25 words. Four different text corpora were selected and statistically analysed. The selection of sentences for the final corpus was based on a two-stage process. In the first stage an analysis based on statistical criteria was performed. In the second stage the final text was chosen based on the results of the first stage /10/.

2.2 Audio recordings

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

Since the speaker was a professional radio news speaker the speech contained no disfluencies (i.e. filled pauses, repetitions and deletions) although for this particular speaker there are some indices for hesitations in the form of pauses and lengthening. Compared to the German corpus /19/ of resembling extent used in /8/, the percentage of hesitations differed essentially (<0,5% German, >15% Slovenian).

2.3 Phonetic transcription

The phonetic transcription was managed using a two step conversion module. The first step was realised with a rule-based algorithm. Subsequent to the first step the second step was designed using 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 those cases where no morphological lexica was available. The first rule based stress assignment was done, followed by a grapheme-to-phoneme conversion procedure.

The step of stress marking before grapheme-to-phoneme conversion is very important for the 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 was used as a data source for training the neural networks /10/.

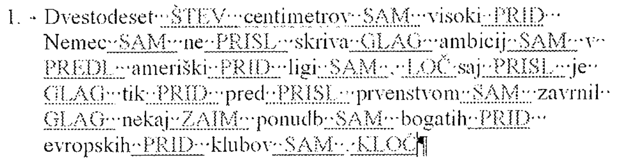
2.4 Part of Speech Tags

The text corpus was hand-labelled using the following part-of-speech tags (POS) /15/:

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, adverbs, ...),
7. PRED for predicative,
8. PREP for prepositions,
9. CONJ for conjunctions,
10. PART for particle,
11. INT for interjection,
12. IPUNC for inter punctuation and
13. EPUNC for end punctuation

All tags were combined in an environment where tracking and correcting tags were simplified for the labellers /13/.

Compared to the tag-set in the German corpus used in /8/, the tag-set for the Slovenian language is smaller. The difference in size occurs because the Slovenian corpus is hand-tagged and no reliable tagger currently exists for a large tag-set (Figure 1).



1. → Dvestodeset··ŠTEV··centimetrov··SAM··visoki··PRID··
Nemec··SAM··ne··PRISL··skriva··GLAG··ambicij··SAM··v··
PREDL··ameriški··PRID··ligi··SAM·,·LOČ·saj··PRISL··je··
GLAG··tik··PRID··pred··PRISL··prvenstvom··SAM··zavrnil··
GLAG··nekaj··ZAIM··ponudb··SAM··bogatih··PRID··
evropskih··PRID··klubov··SAM··.·KLOČ¶

Figure 1: With POS and prosody breaks labeled clause

2.5 Phonetic segmentation and labelling

The spoken corpus was phonetically transcribed using HTK. Along with standard nomenclature two special markers were used for pauses between phonemes. "sil" denotes the silence before and after sentence. "sp" denotes the silence between words in the sentence. Both were determined with one state HMM and all phonemes with three state HMM respectively in the HTK environment (Figure 2).

```
#!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
...
92400000 92950000 k
92950000 93300000 l
93300000 94300000 u:
94300000 94550000 b
94550000 94950000 O
94950000 96500000 W
96500000 100350000 sil
```

Figure 2: An example of phonetic segmented and labeled text

3 Prosody breaks labelling

3.1 Labelling inventory

Since no inventory for symbolic prosody breaks labels is defined for the Slovenian language, we decided to use similar labels to those used in /5/ and in /7/. Thus the prosody break labels were determined through acoustic perceptual sessions and the text was labelled speaker dependent. The following inventory of prosody break labels was used for labelling the corpus /12/:

- 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.

3.2 Semi- automatic prosody breaks labelling

A tool intended to help the labeller (novice or expert) to make decisions about prosody breaks within each sentence was designed /14/. The tool indicates possible prosody boundaries, which depend on the segmented pauses in spoken corpora and pitch accents.

Experiments on multi-lingual databases (3 languages) have shown that the strategy of segmenting the speech signal with pauses, yields a significant improvement in annotation accuracy /18/.

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.

The tool indicates markers for prosody boundaries taking phonetic segmentation of pauses into account. The position of prosody boundaries is selected by considering the duration of silence between words. The decision of indication is made by comparison with a specific threshold. This threshold can be changed manually and can be tuned according to a specific speaker /14/.

4 Labelling results

Labelling experiments were performed for prosody breaks and, additionally, one experiment for accents labelling. In the first experiment prosody breaks labels were marked at positions indicated by the tool.

Additionally, a careful analysis of the f0-contours, energy contours, and perception lead to the insertion at positions that the tool did not indicate. This labelling 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.

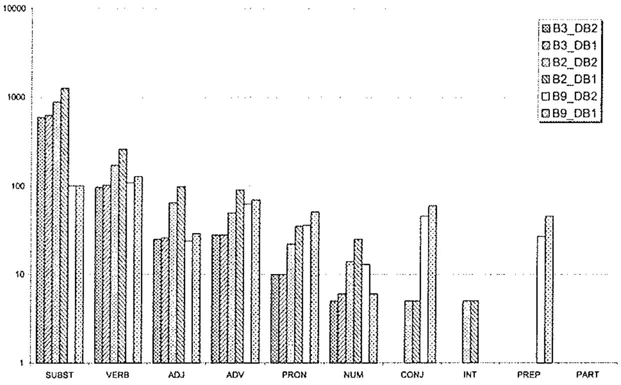


Figure 3: The occurences of prosody break labels in DB1 and DB2

The frequencies of occurrence for each labelled break for each POS tag are presented in Figure 3. The increase of B2 tags in DB1 compared to DB2 is proportional for almost all POS tags. The increase of B9 labels is evidently minor to the increase of B2 labels and in our opinion is strongly speaker dependent.

The complete labelling of 600 clauses had now been performed. It was possible to detect 77,95% of all breaks (over 93 % for B3) and considerable shorten the time needed for labelling the database with the semiautomatic method used /13/.

5 Phrase break prediction module

5.1 Input parameters

Which parameters are relevant for symbolic prosody label prediction remains an open research question. A carefully chosen feature set can help to improve prediction accuracy, however, finding such a feature set is work intensive. In addition linguistic expert knowledge can be necessary and the feature set found can be language and task dependent. A feature set which is commonly used and which seems to be relatively independent of language and task is part-of-speech (POS) sequences /8/.

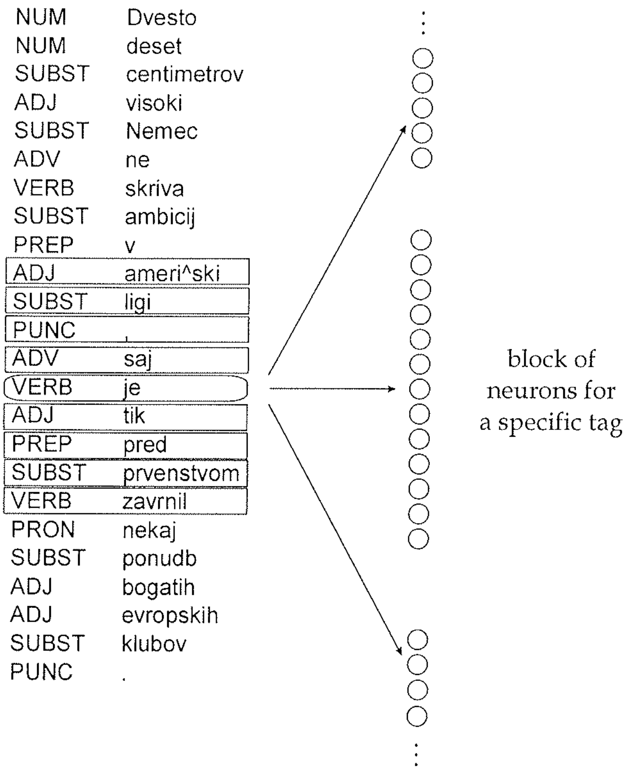


Figure 4: Input mapping into the MLP structure

POS sequences of length four to the left and right of the position in question were used. For the input of our prediction model the POS sequences were coded with a ternary logic (-1 for a non-active node, +1 for an active node, 0: not valid) /4/. Thus for each POS tag a vector was obtained with a dimension of the size of the tag-set. The size of our tag set was 13. Using a POS sequence length of four to the left and right for the Slovenian language, we achieved m = (4+1+4) * 13 = 117 dimensions. The dimension of our input vector as well as tag-set is similar to the English (German) language prediction tests in /8/ where a tag-set of length 14 was used (Figure 4).

5.2 Design of the neural network model

MLP networks are normally applied to performing supervised learning tasks, which involve iterative training methods to adjust the connection weights within the network. This is commonly formulated as a multivariate non-linear optimisation problem over a very high-dimensional space of possible weight configurations./3/. One (two) hidden layers connected to the input layer (and bias) were used and tests were performed with 30-40 neurones in each layer (Figure 5). The output layer was reduced to one (two) single outputs. The results presented (cf. Experiments and results) are for MLP structures with one hidden layer.

5.3 Training and pruning method

A variation of the standard back-propagation algorithm was applied to train the NN - VarioEta /11/. Patterns were selected with a quasi stochastic procedure. An approximation of the gradient was used for the determination of search direction **d**ᵢ, computed from the average over a subset **M** of all patterns:

$$\nabla E^{M}(w)=\frac{1}{|M|}\sum_{n\in M}\nabla E^{n}(w) \qquad (1)$$

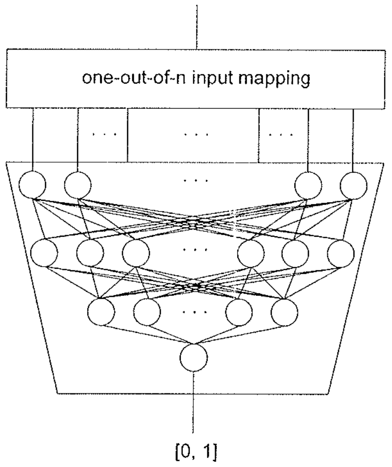


Figure 5: The MLP NN architecture

Note that $\mathbf{d}_i$ is not normally parallel to - ∇M (1) and thus VarioEta does not constitute a pure gradient method. The subset M was determined with a permutation procedure known as "drawing without replacement". During the first adaptive step, X patterns were chosen at random, during the second step X of the remaining |M| (|M| denoted the number of elements of M), and so forth. Each pattern had the same probability of being chosen. After Y steps, every pattern would be read in exactly once. The initial setting for the training started with a learning rate of η =0.05, decreasing its value every 10 epochs. The values for |M| were varied between 50 and 250.

The network was trained so that the output error on validation data converged to a local minimum using the *late stopping strategy* /9/. A careful examination of correlation between the transformed ternary coded input patterns was performed. Nodes with high correlation rates (over 90 %) were removed from the input training inventory. Irrelevant and destructive weights were eliminated by early brain damage /16/. Node pruning based on sensitivity analysis was used to reduce the number of hidden neurones /9/.

6 Experiments and results

After labelling the corpus as described in the preceding section, the two databases (DB1 and DB2) were used to train the phrase break prediction module of our text-to-speech system. For both databases the B9-labels marking hesitations were removed prior to training, since hesitations generally occur at positions where a break seems unsuitable. 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 generalisation set (20% of the data). All results were determined on the independent generalisation set.

Tests were made comparing the prediction results using labelled data in DB1 and DB2 as training data for our phrase break prediction module. The results were significantly better for DB2 (Table 1). 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 cases where additional breaks are labelled based on perception. It also showed that the prediction accuracy degrades only insignificantly for those cases where the 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, this tool can be used for labelling without performance loss. This results in a significant reduction in time needed to label a database.

Table 1. Comparisons of phrase break prediction for DB1 and DB2

breaks	DB1	DB2
minor/major	74,05 %	79,37%
minor=major	77,74 %	80,60 %

Considering the discussed results in previous paragraphs and the fact that tests of prediction accuracy no matter what approach is used for prediction, are preferably made with hand labelled corpora /1/, /8/, for comparison reasons it was decided to conduct further experiments with the DB1.

Table 2. Results for phrase break prediction DB1

breaks	B correct	NB incorrect	overall
DB1	77,74 %	4,95 %	94,03 %

Table 3. Confusion matrix for the generalisation data

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

In tables 2 and 3 the results are presented for the phrase break prediction module. Despite the limited labelling material available (600 labelled clauses were used) and the multi layer perceptron (MLP) NN structure used for prediction, the results accomplished are comparable to prediction accuracy for German /8/ and English /1/.

For the prediction of breaks (B correct) the results are equivalent to the achieved accuracy prediction of B correct (77,67 %) for German in /8/ or nearly equivalent to achieved accuracy prediction of B correct (79,27 %) for English in /1/ despite a much smaller inventory of clauses used. Slightly better overall and non-break (NB incorrect) prediction accuracy was achieved. In the next tables (Table 4, Table 5, Table 6, Table 7) the results for isolated major/minor phrase breaks prediction accuracy are presented.

Table 4. Results for B3 phrase break prediction DB1

breaks	B3 correct	NB3 incorrect	overall
DB1	74,05 %	0,15 %	98,36 %

Table 5. Confusion matrix for the generalisation data on B3

predicted/actual	breaks	non-breaks	all predicted
breaks	605	19	624
non-breaks	212	13211	13423
all actual	817	11432	

The results presented for achieved phrase break predictions of B3 markers are superior to the prediction for B2 markers. The reason for the difference is assumed to be the percentage of additionally hand labelled B2 markers in DB1. The first experiments using the DB2 for input data are much more promising, although the database is not covering the entire inventory of hand labelled phrase breaks in the Slovenian corpus. Therefore some additional experiments (time consuming perceptual sessions) will be necessary to evaluate the suitability of the covered prosody features.

Table 6. Results for B2 phrase break prediction

breaks	B2 correct	NB2 incorrect	Overall
DB1	66,13 %	3,71 %	92,43 %

Table 7. Confusion matrix for the generalisation data on B2

predicted/actual	breaks	non-breaks	all predicted
breaks	1189	454	1643
non-breaks	609	11795	12404
all actual	1798	12249	

7 Conclusion

This paper presents an approach for the labelling and classification of symbolic prosody phrase breaks for the Slovenian language. A universal tool for hand labelling the corpora with prosodic markers was designed and used for the labelling of Slovenian corpus for phrase breaks and accent prediction. This tool can be seen as a first step towards the semi-automatic labelling of prosody features for the Slovenian language. Our aim was to design a tool suitable for performing multilingual prosody labelling. Only features for prosody prediction were, therefore, used which seem to be relatively independent of language and task. Our conclusion is that the approach used - the segmented pauses in the speech corpus for phrase boundary indication - is very useful for the symbolic prosody breaks labelling of the Slovenian language. Firstly, it considerably reduces the time needed for labelling and, secondly, it provides a high level of support to a labeller for consistent labelling of prosodic events. Nevertheless the data obtained with our labelling tool seems to be much more suitable for the training of our prediction module. The NN structure used is accurate enough when compared to other, more complex, NN structures and approaches in prediction of symbolic prosody markers.

The database for the Slovenian language labelled with the proposed tool was used to train our phrase break prediction module /13/. The achieved prediction accuracy marks a good entry-level for phrase break prediction accuracy of the Slovenian language. Nevertheless a minor clause inventory was used, compared to other approaches, with equivalent or superior success in phrase break prediction accuracy.

We also conclude that the simple NN structure proposed is suitable for implementation in DSP environments for speech synthesis as an ad-on prosodic module.

8 References

/1/ Black A. W., Taylor P. (1997). Assigning Phrase Breaks from Part-of-speech Sequences. Proceednings Eurospeech 97, Rhodes, Greece.

/2/ 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.

/3/ Gallagher M. R. Multi-Layer Perceptron Error Surfaces: Visualization, Structure and Modelling. PhD Thesis, University of Queensland, Department of Computer Science and Electrical Engineering, 2000.

/4/ Hain H.-U. (1999). Automation of the training procedure for neural networks performing multi-lingual grapheme to phoneme conversion, Proceedings Eurospeech 99, Budapest, Hungary.

/5/ Kompe R. (1997). Prosody in Speech Understanding Systems. Springer - Verlag Berlin Heidelberg, Lecture Notes in Artificial Inteligence.

/6/ Malfrere F., Dutoit T. and Mertens P. (1998). Fully automatic prosody generator for text-to-speech. ICSLP 98, Sydney Australia.

/7/ Mihelič F., Gros J., Nöth E., Dobrišek S., Žibert J. (2000). Recognition of Selected Prosodic Events in Slovenian Speech, Language Technologies, Ljubljana, Slovenia.

/8/ Müller A. F., Zimmermann H.G., and Neuneier R. (2000). Robust Generation of Symbolic Prosody by a Neural Classifier Based on Autoassociators. Proceednings ICASSP 00, Istanbul, Turkey.

/9/ Neuneier R., Zimmermann H.G. How to train neural networks. In Ohr G. B. and Müller K.-R., editors Neural Networks: Tricks of the Trade. Springer Verlag, Berlin, 1998.

/10/ Rojc M., Kačič Z. (2000). Design of Optimal Slovenian Speech Corpus for use in the concatenative Speech Synthesis System. LREC 00, Athens, Greece.

/11/ Senn Version 3.0 User Manual. SIEMENS AG. 1998

/12/ SI1000 (1998). Prosodic Markers Version 1.0, Bavarian Archive of Speech Signals. University of Munich, Institute of Phonetics, Germany.

/13/ Stergar J. (2000). Determining Symbolic Prosody Features with analysis of Speech Corpora. Master Thesis. University of Maribor. Faculty for EE. and Comp. Sci.

/14/ 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, Slovenia.

/15/ Toporišič J. (1991). Slovenska slovnica. Založba obzorja Maribor.

/16/ Tresp V., Neuneier R., Zimmermann H. G.. Early Brain Damage. In Advances in Neural Information Processing Systems, volume 9. MIT Press, 1997.

/17/ Vereecken H., Martens J. P., Grover C., Fackrell J., Van Coile B. (1998). Automatic prosodic labeling of 6 languages. ICSLP 98, Sydney Australia.

/18/ Vereecken H., Vorstermans A., Martens J. -P. and Van Coile B. (1997). Improving the Phonetic Annotation by means of Prosodic Phrasing. Proceedings Eurospeech 97. Rhodes, Greece.

9 Web References

/19/ Institut für Phonetik und sprachliche Kommunikation: Siemens Synthese Korpus - SI1000P, http://www.phonetik.uni-muenchen.de/Bas/.

mag. Janez Stergar, univ. dipl. inž. el.
red. prof. dr. Bogomir Horvat, univ. dipl. inž. el.
University of Maribor
Faculty of Electrical Engineering and Computer Science
Smetanova ulica 17
2000 Maribor
tel.: +386 2 220 7203
fax.: +386 2 251 1178

Prispelo (Arrived): 27.05.2002 Sprejeto (Accepted): 28.06.2002