Recognition and Labeling of Prosodic Events in Slovenian Speech

France Mihelič[1], Jerneja Gros[1], Elmar Nöth[2], and Volker Warnke[2]

[1] University of Ljubljana, Faculty od Electrical Engineering,
Tržaška 25, SI-1000 Ljubljana, Slovenia
{mihelicf, nejka}@fe.uni-lj.si,
WWW home page: http://luz.fe.uni-lj.si
[2] Universität Erlangen-Nürnberg,
Martensstrasse 3, 91058 Erlangen, BRD
{noeth, warnke}@informatik.uni-erlangen.de

Abstract. The paper describes prosodic annotation procedures of the GOPOLIS Slovenian speech data database and methods for automatic classification of different prosodic events. Several statistical parameters concerning duration and loudness of words, syllables and allophones were computed for the Slovenian language, for the first time on such a large amount of speech data. The evaluation of the annotated data showed a close match between automatically determined syntactic-prosodic boundary marker positions and those obtained by a rule-based approach.

1 Introduction

Research and development in the field of spoken language technology encouraged and supported by the European Community resulted in many experimental and also commercially available systems for speech recognition, understanding, synthesis and dialogue for different European languages [1].

Noticeable progress in this field in the recent years was also made for one of the minor European languages, Slovenian. As a result of our own investigations and our collaboration with foreign partners, a multilingual dialogue system was developed [3]. Slovenian speech recognition, understanding and synthesis systems are the most important achievements of the research group in Slovenia [4].

The usage of prosodic parameters in speech recognition and understanding and quality prosody modeling in speech synthesis resulted in large improvements in the performance of these systems [5–7]. At present not much work has been done in this field for the Slovenian language. The research group at the University Erlangen-Nürenberg has some important research results in the field of prosodic analysis for the German language [5, 9, 8] and also some experience with prosody processing for other languages [10]. In a combined effort of the Slovenian and the German research groups procedures for automatic prosodic parameter measurement and evaluation were tested on Slovenian speech material. We tried to recognize prosodic events and evaluate them statistically.

2 Selection of prosodic events

Syntactic-prosodic boundaries along the lines of [5], annotated for transliterations of read speech and acoustic-prosodic boundaries and word accents labeled via acoustic perceptual sessions were chosen for the initial experiments.

- M3: clause boundaries
- M2: constituent boundaries likely to be marked prosodically
- M1: boundaries that syntactically belong to the normal constituent boundaries probably not marked prosodically because they are "close" to a M3 boundary
- M0: every other word boundary

Table 1. Syntactic-prosodic boundary labels

We performed automatic labeling with syntactic-prosodic boundaries of the Slovenian GOPOLIS speech corpus (50 speakers, 8.645 utterances, 5.077 different corpus sentences) according to [8] pp. 140–144. In the labelling procedure, we distinguished between 4 types of boundaries as listed in Table 1. Here is example of labeled text[1]:

Ali M1 *imate* M2 *kakšno letalo* M3 *čez tri tedne?*
(Is there a flight in three weeks?)

Acoustic-prosodic boundaries and word accent labels were defined as in VERBMOBIL [8]. They are described in Tables 2 and 3.

- B3: prosodic clause boundary
- B2: prosodic phrase boundary
- B9: irregular boundary, usually hesitation lengthening
- B0: every other word boundary

Table 2. Acoustic-prosodic boundary labels

- PA: the most prominent (primary) accent within a prosodic clause
- SA: all other accented words are marked as carrying secondary accent
- UA: unaccented words

Table 3. Word accent labels

Here is a labeled sample sentence[2]:

[1] M0 labels are not indicated in the example.
[2] The default classes B0 and UA labels are not indicated in the example.

Lahko ponovite SA *odgovor* PA , B3 *prosim* PA *?*
(Can you repeat the answer, please?)

3 Data preparation

The experiments on prosodic events classification we wanted to perform required some specific data preparation. Speech signals of read Slovenian texts from the GOPOLIS speech corpus [11] were used for the experiments. The GOPOLIS pronunciation dictionary was extended with stress and syllable markers. Special categories denoting *silence, non-word, consonants* and *syllable-root* were added. A special format for representing word graphs was used.

The segmentation of the speech signals was performed automatically using the ISADORA net [12]. An important feature of the Isadora net environment is the automatic detection of silence segments not indicated in the transliteration.

The text corpus of the GOPOLIS database has been created automatically using a context-free grammar consisting of 189 sentence templates [15]. The sentence templates cover the most frequent dialogue situations occurring at airline timetable information retrieval.

In order to obtain a version of the GOPOLIS database annotated with boundary information, M1, M2 and M3 labels were inserted into the sentence templates. Then the sentence generation process was repeated.

The M2 markers were generally set according to rhythmic constraints. M1 markers — as described above — were set, if an M2 boundary was too close to the beginning or end of the utterance or too close to an M3 boundary, i.e. they were also set according to rhythmic constraints. M3 markers were placed between main/subordinate clauses and around embedded clauses.

Acoustic-prosodic boundaries and word accent labels were marked manually by perception tests using visualisation of speech signals and listening. In total, 1000 signals (6 speakers) were labeled, representing 12% of the whole GOPOLIS speech corpus. The annotation was conducted by the first author; intralabeller consistency has not yet been evaluated.

4 Features selection and pattern classification

Classification using only prosodic feature sets describing duration segmental characteristics on the word level, speaking rate, energy and pitch was performed. Pitch periods were computed for the entire GOPOLIS database based on the inverse filtering algorithm [9]. Various statistical parameters previously determined on the training set were used in the duration and energy normalization procedures [9] (pp. 165–182). All 95 features forming "the best word features set" according to experiments performed on German speech [9] (page 258) were computed for the Slovenian speech data.

Energy and pitch features are based on the short-term energy and F0 contour, respectively. Some of the features that are used to describe a pitch contour are shown in Figure 1. Additionally, we use the mean and the median as features [2].

Neural nets were used for classification. The SNNS software [13] was used for neural net modeling. Several net configurations and initializations were tested in the experiments.

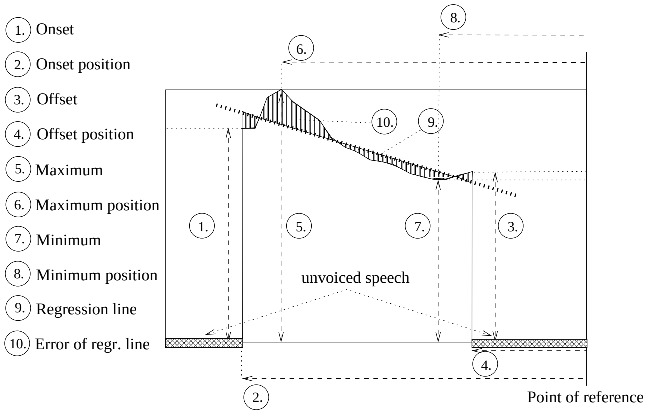

Fig. 1. Example of features used to describe a pitch contour.

5 Recognition results

Four classification schemes with different clustering for the syntactic-prosodic boundaries were tested. The results indicating overall recognition rate and class-wise computed recognition rate are displayed in Table 4. 5 female and 5 male speakers were selected for the test set while the remaining 40 speakers were used in the training set.

The classification of M3 markers versus other categories and classification into three classes (M3 : M2 : M1, M0) are the most important for further use in semantic parsing of recognized word strings [6]. By comparing the obtained results to the results obtained on the ERBA corpus [14] (82%, 72%), the results on Slovenian speech data are even slightly better. This indicates at least the correctness of the automatic procedure for boundary labeling and also the appropriate feature selection and classification procedure for the Slovenian language.

Due to the lack of manually labeled data[3] the results of the experiments on acoustic-prosodic boundary classification and word accent classification are not so reliable. They might be significantly improved using a larger training set. The results are displayed in Table 5. Two turns of classification experiments were performed. Both times 5 speakers were used for training and 1 speaker for testing.

[3] Only approx. 12% of all GOPOLIS corpus was manually labeled.

Classification schemes	Recognition overall	Recognition classwise
M0 : M1 : M2 : M3	74.1%	77.5%
M0 M1 : M2 : M3	80.7%	80.3%
M0 M1 M2 : M3	87.3%	87.7%
M0 : M1 M2 M3	87.5%	87.6%

Table 4. Recognition results for different classification schemes for syntactic-prosodic boundaries. App. 40000 labels were used in the training set, 8400 in the test set, label M3 at the end of the utterance was not included in the tests.

Speaker	Classification schemes	Recogn. overall	Recogn. classwise
01M	B0 : B9 : B2 : B3	73.9%	79.8%
03M	B0 : B9 : B2 : B3	66.8%	67.2%
01M	UA : SA : PA	64.1%	62.5%
03M	UA : SA : PA	72.5%	69.9%

Table 5. Recognition results for different classification schemes for acoustic-prosodic boundaries and word accent labels. Approx. 4500 labels were used in the training set and 840 labels in the test set for acoustic-prosodic boundaries; the label B3 at the end of the utterance was not included in the tests. Approx. 5000 labels in the training set and 1000 labels in the test set for the word accent labels.

6 Conclusion

Useful knowledge on prosodic events for the Slovenian speech was gained as an intermediate results of the described experiments. Pitch periods for the complete GOPOLIS speech data base were computed and they can be used as a reference for further investigations. Several statistical parameters concerning duration and loudness of words, syllables and allophones were computed for the Slovenian language, for the first time on such a large speech database.

The obtained research results encourage us to continue the co-operation of both partner groups. Data collection and annotation as well as development of semantic parsers, using prosodic information as additional input, will be the next steps of our research. The computed prosodic parameters will also be applied in the prosody prediction process for Slovenian text-to-speech synthesis.

7 Acknowledgement

France Mihelič thanks the DAAD for its support enabling his study visit in Erlangen from May to July 1999, during which most of the presented work was done. Special thanks go to F. Gallwitz and J. Buckow who helped organizing the experiments in Erlangen and to S. Dobrišek and J. Žibert from the University of Ljubljana for their support.

References

1. Andersen, P.: Language Technology and Multilinguality - The European Dimension. Invited Lecture, Proceedings of the Conference Language Technologies for the Slovene Language, Eds. T. Erjavec, J. Gros, Ljubljana (1998) 9–13
2. Buckow, J., V. Warnke, R. Huber, A. Batliner, E. Nöth , H. Niemann: Fast and Robust Features for Prosodic Classification. Proc. Text, Speech and Dialoge, Lecture Notes in Artificial Intelligence 1692, Marienbad (1999) 193–198
3. Aretoulaki, M, S. Harbeck, F. Gallwitz, E. Nöth, H. Niemann, J. Ivanecky, I. Ipšić, N. Pavešić, V. Matoušek: SQEL: A Multilingual and Multifunctional Dialogue System. Proc. Int. Conf. on Spoken Language Processing, (1998) 855–858
4. Ipšić, I, F. Mihelič, S. Dobrišek, J. Gros, N. Pavešić: A Slovenian Spoken Dialog System for Air Flight Inquires. Proceedings of the Eurospeech'99, Budapest, Hungary (1999) 2659–2662
5. Batliner, A., R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth: M = Syntax + Prosody: A syntactic-prosodic labelling scheme for large spontaneous speech databases. Speech Communication, **25** (1998) 193–222
6. Boros, M., J. Haas, V. Warnke, E. Nöth, and H. Niemann: How Statistics and Prosody can guide a Chunky Parser. Proc. of the AIII Workshop on Artificial Intelligence in Industry, Stara Lesna, Slovakia (1998) 388–398
7. Gros, J., F. Mihelič, N. Pavešić: Speech Quality Evaluation in Slovenian TTS. First International Conference on Language Resources and Evaluation, Eds. A. Rubio, N. Gallardo, R. Castro, A. Tejada, Vol. I, Granada, Spain (1998) 651–654
8. Kompe, R.: Prosody in Speech Understanding Systems. Springer-Verlag Berlin Heidelberg, Lecture Notes in Artificial Intelligence 1307 (1997)
9. Kießling A.: Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Schaker Verlag (1997)
10. Haas, J., V. Warnke, H. Niemann, M. Cettolo, A. Corazza, D. Falavigna, G. Lazzari: Semantic Boundaries in Multiple Languages. Proceedings of the Eurospeech'99, Vol. I, Budapest, Hungary (1999) 535–538
11. Dobrišek, S., J. Gros, F. Mihelič, N. Pavešić, . Recording and Labelling of the GOPOLIS Slovenian Speech Database. First International Conference on Language Resources and Evaluation, edits. A. Rubio, N. Gallardo, R. Castro, A. Tejada, Vol. II, Granada, Spain (1998) 1089–1096
12. Schukat-Talamazzini, E.G.: Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen. Vieweg, Braunschweig (1995)
13. Stuttgart Neural Network Simulator. http://www-ra.informatik.uni-tuebingen.de/SNNS (1999)
14. Bakenecker, G., U. Block, A. Batliner, R. Kompe, E. Nöth, P. Regel-Brietzmann: Improving Parsing by Incorporating 'Prosodic Clause Boundaries' into a Grammar. Proc. Int. Conf. on Spoken Language Processing, Vol. 3, Yokohama (1994) 1115–1118
15. Gros, J., F. Mihelič, N. Pavešić: Sentence hypothesisation using Ng-gram models. Proceedings of the Eurospeech'95, Vol. 3, Madrid (1995) 1759–1762