Recognition of Selected Prosodic Events in Slovenian Speech

France Mihelič*, Jerneja Gros*, Elmar Nöth†, Simon Dobrišek* and Janez Žibert*

*Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, 1001 Ljubljana, Slovenia
{mihelicf, nejka, simond, janezz}@fe.uni-lj.si

† Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen-Nürnberg
Martensstrasse 3, 91058 Erlangen, BRD
noeth@informatik.uni-erlangen.de

Abstract

Prosodic annotation procedures of the GOPOLIS Slovenian speech data database and methods for automatic classification of different prosodic events are described in the paper. Several statistical parameters for duration and loudness of words, syllables and allophones were computed for the Slovenian language, for the first time on such a large amount of Slovene speech data. The evaluation of the annotated data shows 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 technologies 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 (Andersen, 1998). The first steps towards multilingual speech technology systems were also performed. In Germany, for instance, collaboration between many research partners in the Verbmobil project (Čavar et al., 1998) was established.

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 (Aretoulaki et al., 1998). Slovenian speech recognition, understanding and synthesis systems are the most important achievements of the research group of Slovenian scientists in the Laboratory of artificial perception, Faculty of electrical engineering at the University of Ljubljana (Ipšić et al., 1999).

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 (Batliner et al., 1998; Boros et al., 1998; Gros et al., 1998). 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 (Batliner et al., 1998; Kißling, 1997; Kompe, 1997) and also some experience with prosody processing for other languages (Haas et al., 1999). 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 (Batliner et al., 1998), 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 as M2 but are most certainly not marked prosodically because they are "close" to a M3 boundary
- M0: every other word boundary

Table 1: Syntactic-prosodic boundary labels

For the training of statistical classifiers large amounts of labeled training data are needed. Prosodic labeling on the basis of perception tests is very time consuming. Furthermore, it does not exactly reflect what is needed during the syntactic analysis of speech. Therefore, 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 (Kompe, 1997) pp. 140–144.

In the labelling procedure, we distinguished between 4 types of boundaries as listed in Table 1. Here are some examples of labeled text[1]:

Lahko M1 *ponovite odgovor* M3 *prosim?*
(Can you repeat the answer please?)
Ali M1 *imate* M2 *kakšno letalo* M3 *čez tri tedne?*
(Is there a flight in three weeks?)
Letite M2 *na relaciji* M3 *Helsinki* M3 *Zuerich?*
(Do you fly the route Helsinki Zuerich?)

Acoustic-prosodic boundaries and word accent labels were defined as in VERBMOBIL (Kompe, 1997). They are described in Tables 2 and 3.
Here is a sample list of sentences labeled with acoustic-prosodic boundaries and word accents [2]:

[1] M0 labels are not indicated in the examples.

[2] The default classes B0 and UA labels are not indicated in the

- 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

Ja PA, <pause> B3 *res* PA *!*
(Yes, really!)
Lahko ponovite SA *odgovor* PA , B3 *prosim* PA *?*
(Can you repeat the answer, please?)
Bi bila SA *mogoče* SA *kakšna* SA *direktna* PA *letalska* SA *veza* B3 *čez tri* PA *dneve* SA *?*
(Is there a direct connection in three days?)

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 (Dobrišek et al., 1998) 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. Table 4 displays an example of a word–graph.

The segmentation of the speech signals was performed automatically using the ISADORA net (Schukat-Talamazzini, 1995), based on HMM acoustic speech modeling and transliterated texts. An important feature of the Isadora net environment is the automatic detection of silence segments not indicated in the transliteration.

3.1. Automatic syntactic–prosodic boundaries determination

The text corpus of the GOPOLIS database has been created automatically using a context-free grammar consisting of 189 sentence templates (Gros et al., 1995). The sentence templates cover the most frequent dialogue situations occurring at airline timetable information retrieval. The templates were created after tedious listening, transcribing and analyzing of 15 hrs of recordings of real situation dialogues between anonymous clients and telephone operators at the Adria Airways information center. 22,500 different sentences for short introductory inquiries, long inquiries and short confirmations were produced using the sentence templates. 5,077 of them were randomly chosen to form the final GOPOLIS sentence corpus.

examples.

```
BEGIN_LATTICE
```

1	2	[−]	1.00	2	18	(AP (Z - 2))
2	3	*ne*	1.00	19	55	(AP (Z n 19 E: 32))
3	4	[−]	1.00	56	80	(AP (Z - 56))
4	5	*to*	1.00	81	94	(AP (Z t 81 o: 87))
5	6	*pa*	1.00	95	115	(AP (Z p 95 a 105))
6	7	*ne*	1.00	116	132	(AP (Z n 116 E: 120))
7	8	[−]	1.00	133	156	(AP (Z - 133))

```
END_LATTICE
```

Table 4: Word-graph for the sentence "Ne, to pa ne!" (meaning "No, not this!"). Col. 3: description of the speech segment (word, pause), Col. 5: speech segment start time, col. 6: speech segment end time, Col. 7: word allophones and their start time. The time unit corresponds to 10 msec.

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. For example, the DEP_CITY ARR_CITY category combination indicating the flight arrival–destination city was always delimited by a M3 marker since we assumed that the speakers would pronounce the two city names separately as they convey essential semantic information. Similarly, time and date expressions were separated from the rest of the text using M3 markers. Since one sentence template can produce sentences of various length, many of the M3 markers could not be predicted automatically, esp. the breathing pauses.

3.2. Acoustic-prosodic boundaries and word accent labels determination

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 (Kißling, 1997). Various

statistical parameters previously determined on the training set were used in the duration and energy normalization procedures (Kißling, 1997) (pp. 165–182). All 95 features forming "the best word features set" according to experiments performed on German speech (Kißling, 1997) (page 258) were computed for the Slovenian speech data.

The aim of the extraction of prosodic features is to compactly describe the properties of the speech signal which are relevant for the detection of prosodic events. Prosodic events, such as phrase boundaries and phrase accents, manifest themselves in variations of speaking-rate, energy, pitch, and pausing. The exact interrelation of these prosodic attributes is very complex. Thus, our approach is to find features that describe the attributes as exactly but also as compactly as possible.

At each edge of the WHG, not only the current edge (i.e. the current word interval) is used for feature extraction but also intervals containing several words. These intervals from the beginning of word f to the end of word t are referred to by $I_{(f,t)}$. Intervals that we use are e.g. $I_{(-2,-1)}$ or $I_{(-1,0)}$. At the end of the word *"not"* in the utterance shown in Figure 1 the Interval $I_{(-2,-1)}$ e.g. denotes the time interval from the beginning of the word *"Of"* to the end of the word *"course"*.

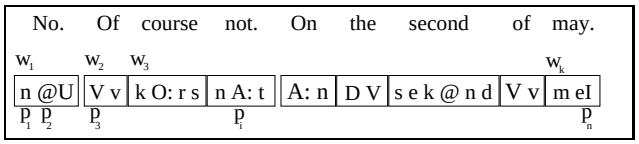

Figure 1: Utterance *"No. Of course not. On the second of May."* with the phoneme sequence in SAMPA notation.

Each of the features that were used in the experiments corresponds to an interval as described above. The pause features are easily extracted: These are simply the duration of *filled pauses* (e.g. "uhm", "uh", ...) and *silent pauses*. Energy and pitch features are based on the short term energy and F0 contours. Duration features should capture variations in speaking-rate and are based on the duration of speech units. A normalization of energy, duration, and pitch features can be performed in order to take phone intrinsic variations and the optional use of prosodic marking into account.

As mentioned above, 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 2. Additionally, we use the mean and the median as features (Buckow et al., 1999).

Neural nets were used for classification. The SNNS software (SNNS, 1999) was used for neural net modeling. Several net configurations and initializations were tested in the experiments.

5. Recognition results

Four classification schemes with different clustering for the syntactic-prosodic boundaries were tested. The results indicating overall recognition rate and classwise computed

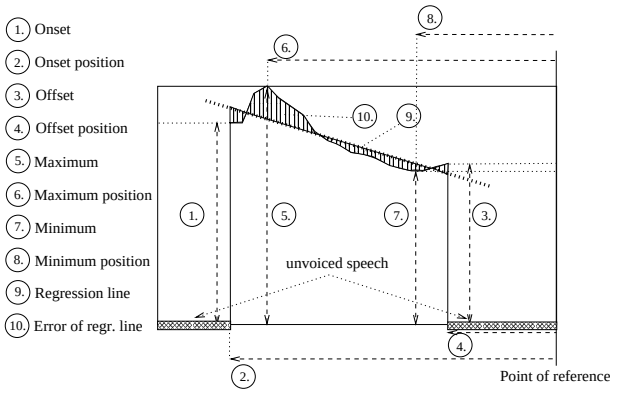

Figure 2: Example of features used to describe a pitch contour.

recognition rate are displayed in Table 5. 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 (Boros et al., 1998). By comparing the obtained results to the results obtained on the ERBA corpus (Bakenecker et al., 1994) (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.

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 5: 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.

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 6.

Two turns of classification experiments were performed. Both times 5 speakers were used for training and 1 speaker for testing.

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

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

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 6: 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.

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.

It can be seen that two languages belonging to different language groups – Germanic and Slavonic – could successfully be processed using the same syntactic and prosodic modelling procedure. In the near future, there will still be more resources available for the major European languages than for the minor ones. It seems to be likely that this disadvantage can be by–passed at least partly by approaches like those described in this paper.

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 W. Warnke, F. Gallwitz and J. Buckow who helped organizing the experiments in Erlangen.

8. References

Andersen, P., 1998. 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, 9–13.

Buckow, J., V. Warnke, R. Huber, A. Batliner, E. Nöth , H. Niemann, 1999. Fast and Robust Features for Prosodic Classification. *Proc. Text, Speech and Dialoge*, Lecture Notes in Artificial Intelligence 1692, Marienbad, 193–198.

Čavar, D., W. Menzel, 1998. VERMOBIL: A Speech-to-Speech Translation System. Introductory Lecture, *Proceedings of the Conference Language Technologies for the Slovene Language*, Edts. T. Erjavec, J. Gros, Ljubljana, 25–28.

Aretoulaki, M, S. Harbeck, F. Gallwitz, E. Nöth, H. Niemann, J. Ivanecky, I. Ipšić, N. Pavešić, V. Matoušek, 1998. SQEL: A Multilingual and Multifunctional Dialogue System. *Proc. Int. Conf. on Spoken Language Processing*, 855–858.

Ipšić, I, F. Mihelič, S. Dobrišek, J. Gros, N. Pavešić, 1999. A Slovenian Spoken Dialog System for Air Flight Inquires. *Proceedings of the Eurospeech'99*, Budapest, Hungary, 2659–2662.

Batliner, A., R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, 1998. M = Syntax + Prosody: A syntactic-prosodic labelling scheme for large spontaneous speech databases. *Speech Communication, 25*, 193–222.

Boros, M., J. Haas, V. Warnke, E. Nöth, and H. Niemann, 1998. How Statistics and Prosody can guide a Chunky Parser. *Proc. of the AIII Workshop on Artificial Intelligence in Industry*, Stara Lesna, Slovakia, 388–398.

Gros, J., F. Mihelič, N. Pavešić, 1998. 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, 651–654.

Kompe, R., 1997. *Prosody in Speech Understanding Systems*. Springer-Verlag Berlin Heidelberg, Lecture Notes in Artificial Intelligence 1307.

Kießling A., 1997. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Schaker Verlag 1997.

Haas, J., V. Warnke, H. Niemann, M. Cettolo, A. Corazza, D. Falavigna, G. Lazzari, 1999. Semantic Boundaries in Multiple Languages. *Proceedings of the Eurospeech'99*, Vol. I, Budapest, Hungary, 535–538.

Dobrišek, S., J. Gros, F. Mihelič, N. Pavešić, 1998. 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, 1089–1096.

Schukat-Talamazzini, E.G., 1995. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen.* Vieweg, Braunschweig.

Stuttgart Neural Network Simulator, 1999. http://www-ra.informatik.uni-tuebingen.de/SNNS/

Bakenecker, G., U. Block, A. Batliner, R. Kompe, E. Nöth, P. Regel-Brietzmann, 1994. Improving Parsing by Incorporating 'Prosodic Clause Boundaries' into a Grammar. *Proc. Int. Conf. on Spoken Language Processing*, Vol. 3, Yokohama, 1115–1118.

Gros, J., F. Mihelič, N. Pavešić, 1995. Sentence hypothesisation using Ng-gram models. *Proceedings of the Eurospeech'95*, Vol. 3, Madrid, 1759–1762.