SloParl - Slovenian Parliamentary speech and text corpus for large vocabulary continuous speech recognition

Andrej Žgank, Tomaž Rotovnik, Matej Grašič,
Marko Kos, Damjan Vlaj and Zdravko Kačič

Laboratory for Digital Signal Processing,
University of Maribor, Smetanova ul. 17, SI-2000 Maribor, Slovenia
andrej.zgank@uni-mb.si

Abstract

This paper present a novel Slovenian language resource - SloParl database. It consists from debates acquired in the Slovenian Parliament. The main goal of the project was to cost-effectly collect a new Slovenian language resource that could be used to augment the available Slovenian speech corpora for developing a large vocabulary continuous speech recognition system. The SloParl speech corpus has a total length of 100 hours. The selected sessions between years 2000 - 2005 were incorporated in it. This speech corpus will be used for lightly supervised or unsupervised acoustic models' training. In accordance with this, the accompanying transcriptions were prepared. The second part of the SloParl database is the text corpus, which covers text of all debates from period 1996 - 2005. It consists of 23M words. It will be used to create different types of speech recogniser's language models. Comparison with other Slovenian language resources showed that SloParl database adds new aspects to the modelling of Slovenian language.

Index Terms: speech database, text corpus, Slovenian, parliamentary debates, speech recognition.

1. Introduction

In the last decade the development of large vocabulary continuous speech recognisers moved from well represented major world's languages (English, Spanish, Mandarine, French, German,...) to less represented ones. One group of such languages are Slavic languages [1, 2, 3, 4], to which also Slovenian belongs to. From LVCSR's point of view is the Slovenian language a complex language, due to its inflectional form, dual and relatively free word order in a sentence. The fact that approximately 10 time larger Slovenian speech recogniser's vocabulary is needed to obtain the same out-of-vocabulary rate as for English language [4], nicely shows the dimension of a Slovenian LVCSR task.

To be able to train a state of the art LVCSR system large amounts of transcribed data are needed for estimating the parameters of acoustic models. There are already some Slovenian LVCSR speech and text resources available that could be used for this task. The oldest of them is the SNABI speech database [5], which can be mainly used for some domain specific scenarios. The Slovenian speech recognition for unrestricted domain can be developed using the BNSI Broadcast News database [6, 7]. A special case among the Broadcast News databases is the SINOD speech database [8], which can be used for Slovenian non-native speech recognition task. Although, these databases are being applied for the LVCSR system development, their size still isn't comparable with the size of speech resources for other major languages. The speech database acquisition and transcription work are, however, tedious and expensive, which represents the main obstacle to build more Slovenian speech resources. Several authors [9, 10, 11, 12] successfully proposed different lightly supervised or unsupervised methods for training the acoustic models, where almost no transcriptions were needed. Such an approach can be used to simplify the speech recogniser's development process in the data preparatory phase. Fortunately, a lot of audio sources are nowadays available over air or Internet that can be used for such creation of speech corpora.

In this paper a novel Slovenian language resource SloParl database will be presented[1]. It consists from speech and text corpus with debates from the Slovenian Parliament[2]. The project was established as cooperation between University of Maribor and Slovenian Parliament, and started at the end of year 2005. The first motivation behind the SloParl project was, to cost-effectly collect additional Slovenian language resources that could be used to improve the development of a Slovenian LVCSR system. It is anticipated, to use the speech corpus for lightly supervised or unsupervised acoustic models training procedure. The text corpus will be used to interpolated parliament's debates language model with a newspaper language model. The second motivation was to create a Slovenian language resource that is comparable with other parliamentary language resources that are currently being used in research community for developing state of the art LVCSR systems [13, 14]. Such systems are being widely applied for different challenging tasks like speech to speech translation, multimodal information retrieval applications,...

The remainder of this paper is organized as follows. The next section describes the organizational aspects and the way how the raw data were acquired. The speech corpus intended to be used for acoustic models training is presented in Section 3, which is followed by a description of the text corpus with parliamentary debates in Section 4. The conclusion and directives for future work are given in the last – fifth – section.

[1] This work was partly financially supported by the Slovenian Research Agency under contract number P2-0069.

[2] http://www.dz-rs.si

2. Organizational aspects and data acquisition

There are two different types of plenary sessions held in the Slovenian Parliament. Regular plenary sessions usually last for four to six days per month. Upon any urgent matter, special sessions are held. There are 90 members of Slovenian Parliament, elected for the duration of four years. The plenary sessions are chaired by a president or a vice-president. Beside the members, also different other speakers appear in debates. They are mainly members of Government, answering members' questions or introducing law's proposals to parliament's members. The debates are in principle an interchange of speaker's turns between the president and all other appearing speakers.

The Parliament's TV is responsible for taking the recordings of plenary sessions, that are then given away to other TV stations. Its is also possible to follow the program of Parliament TV on some cable networks and over the Real video server on parliament's homepage. Due to cooperation with the Slovenian Parliament, the access to archive was granted, which ease our job. The Parliaments TV records each session in parallel on DVD media and on professional analog audio tape recorder. As a lossy audio codec is being used in case of DVD recordings, analog audio tape recordings were chosen as source of raw speech material for the SloParl database. The analog speech signal was captured directly to a personal computer, using a high resolution SoundBlaster Audigy sound card. The raw speech was digitalized with 16-bit quantization and 48kHz sampling rate. Later on, the captured speech signal was downsampled to 16-bit, 16kHz sampling rate. Such speech format proves sufficient quality for a LVCSR systems and it is also compatible with other Slovenian speech resources. During recording sessions some problems regarding the recording level occurred, as different speakers have different speaking style. Usually, one of the loudest speakers was the president, which was therefore used as reference for the calibration of the recording level. The acoustic environment during the parliament's debates can be sometimes very difficult from speech recogniser's point of view, as there is a lot of background speech and noise. Also the large parliament's hall can introduce some echo that can influence the speech recogniser's performance.

For the SloParl speech corpus, recording from period 2000 - 2005 were taken. Here, a special caution was given, to select those plenary sessions, that overlap with news broadcasts included in the Slovenian BNSI Broadcast News database [7]. This cross-section of speech material can be very valuable to narrow the speech recogniser's domain in the case of acoustic and language modeling.

The text of parliament's debates is freely available on the Parliament's homepage. Covered are all parliament's sessions from year 1996 on. The text is the representation of the speeches held during the debates in parliament. It is in the HTML format and was used as raw material for our text corpus. This text was generated in the Parliament's transcriptions office. The main area of text corpus usage will be language modeling in the LVCSR system development process.

3. Speech corpus

The SloParl speech corpus consists of 100 hours of recordings. The majority of this speech material – 92 hours – will be used for training the speech recogniser's acoustic models. The first remaining 4 hours will be used for the system's tuning and development (one regular session from June 2001), while the second 4 hours will be used for LVCSR system's evaluation (one regular session from February 2002). Some basic statistics regarding the speech corpus are given in Table 1.

Table 1: *Statistics of the SloParl speech corpus.*

Parameter	Value
Number of sessions	20
Regular sessions	13
Special sessions	7
Average length (train set)	5:05

The SloParl speech corpus consists of 20 different sessions, where 13 of them were regular and 7 special. The average plenary session in the training set lasts for 5:05 hours. Usually, each session was interrupted with pauses that were excluded from the recordings. There are some differences present between regular and special plenary sessions. Regular sessions included in the training set are in average approximately 12% longer than special sessions. The first ones were scheduled for a long time ahead and are dealing with broad spectrum of political topics. On the other side are special plenary sessions that are organized a short time ahead and cope with some urgent matters. Their topic is as such much more homogenous and therefore simpler for speech recognition, as it is easier to adapt the recognisers vocabulary to narrow topic. Consequently a low out-of-vocabulary rate for speech recognition of inflectional languages can be assured.

An important part of speech database are transcriptions. Here, their production is connected with the overall goal to build a language resource in a cost-effective way. As raw transcriptions, the text from the parliament's homepage was taken. This text is in great extent the correct representation of what was spoken during the debate. Some effects of the spontaneous speech (e.g. fillers, restarts, hesitations,...) are missing, while are on the other side some meta informations (e.g. speaker's name, voting results, time breaks) added directly to text, which have no direct connection with the spoken utterances. The speaker's name, date, number and type of session were kept in transcription header as meta informations, while all the rest was excluded from the transcriptions as overhead.

The training set text was divided into two parts of equal size. Text present in the first part was left as it was at this processing step. For the second part, the time information at speaker turns (beginning and end of each particular turn) will be manually added to the transcriptions. Such time information proved to somewhat improve the LVCSR results [11], and can be added to transcriptions relatively easily and cost-effectively.

To be able to apply the development and evaluation corpus for usage with a LVCSR system, complete transcriptions of spoken material are needed. Therefore, both these database sets will be manually transcribed using the same three phase approach as it was applied for the Slovenian BNSI Broadcast News speech database [6]. The text of debates will be used as initial version of transcriptions. All necessary transcription rules and the working environment (i.e. Transcriber[15]) will be adopted from the Slovenian BNSI Broadcast News project[6].

The transcriptions of speech corpus were analyzed to show the complexity of the SloParl database. Extracted statistics are given in Table 2.

Table 2: *The SloParl speech corpus transcription analysis.*

Parameter	Value
Speaker turns	3665
Number of speakers	255
Number of words	655k
Distinct words	37k

The 20 sessions included in the SloParl speech database consist from 3665 speakers turns. There were altogether 255 different speakers present in the speech corpus. The SloParl speech corpus has 655k words in 100 hours of speech, where 37k of them are distinct. For comparison, 36 hours of BNSI Broadcast News database [7] comprise 1565 speakers who spoke 268k words (37k distinct). Approximately the same number of distinct words in the SloParl and the BNSI database was probably caused by the fact that the topic of parliament debates is narrower than in the case of news broadcasts. As politicians often appear on TV and because political debates are very frequently part of news broadcasts, comparison with the Slovenian BNSI Broadcast News database was performed. Overall 89 speakers appear in both speech corpora, which represents 34.9% of the SloParl corpus. To measure the overlap between content of both speech corpora, the overlap between the SloParl and the BNSI vocabulary was analyzed – it is 46.3%. This low level of overlap shows that the SloParl database successfully extended the Slovenian LVCSR's speech resources.

4. Text corpus

The SloParl text corpus will be in the first place applied for training the LVCSR's language models. The quality of incorporated language models is of extreme importancy for speech recognition of inflectional languages. The text of parliamentary debates is available on parliament's homepage, which was our source of raw text material. All regular and special plenary session from year 1996 onwards are covered. Missing are sessions for the time span 1991-1996. All session between the year 1996 and 2005 were included in the SloParl text corpus.

The text of debates on the homepage employed the UTF-8 encoding, which was converted to ISO 8859-2 encoding. This step was necessary for compatibility reasons with other Slovenian language resources. As it was already mentioned, the raw text material also comprehend some additional informations, like speaker's name, voting results, time brakes... The analysis, which of them could be useful during language model's training, pointed out the following parameters:

- speaker's name,
- date,
- number of session,
- type of session.

This information was kept in the SloParl text corpus as separate meta-data. It is anticipated to use this information for example to cluster different types of sessions (special sessions have homogenous contents) or to interpolate parliament text corpus with newspaper text corpus regarding the dates. All other informations that weren't directly representing the spoken material were excluded from the SloParl text corpus.

Table 3: *The SloParl text corpus statistics.*

Parameter	Value
Years	10
Number of sessions	188
Regular sessions	69
Special sessions	119
Number of debates	781
Number of words	23M
Distinct words	182k

To be able to present the SloParl text corpus, its analysis was performed – the statistics are given in Table 3.

Ten years of parliamentary debates are included in the SloParl text corpus (see Table 3). In this period, 188 sessions were hold, where 69 of them were regular and the rest of them (119) were special. As one session usually takes place on several days, the total number of debates is much higher. In the case of the SloParl database 781 debates were included in the corpus. The SloParl text corpus consists from 23M words, where 182k of them were distinct. The complete plenary sessions that were included in the speech corpus' development and evaluation set were excluded from the text corpus to assure the independency of all sets.

The final step in the presentation of the SloParl database is comparison of text corpus with other Slovenian text resources. An important fact that needs to be emphasized is that the SloParl text corpus incorporate a fair amount of spoken text which is significantly different from the written newspaper text. This is of extreme importancy in the case of developing a LVCSR system, where similarity between acoustic and language models plays an important role. Parliament debates contain both versions of speech: read and spontaneous. The first one can be for example found in the case when a proposed law is being presented. The second one is usually connected with members' comments. The Večer newspaper text corpus (years 1998-2003) that was frequently used for Slovenian speech recognition consists from 105M words, where 660k of them are different. The relation between number of words and number of distinct words in both text corpora is similar, which shows approximately the same level of complexity.

5. Conclusion and future work

The paper presented a Slovenian Parliamentary language resource SloParl, which consists from speech and text corpus. Its generation was a cost effective way to increase the amount of Slovenian language resources that can be used for large vocabulary continuous speech recognition tasks.

The SloParl speech corpus will be used for lightly supervised or unsupervised training of acoustic models. The text corpus will be used to train language models in combination with the newspaper text. The idea is to increase the quantity of available speech material. As a longterm goal, the development of system for automatic transcription of parliamentary debates is planned.

6. Acknowledgements

Authors would like to thank the staff of Slovenian Parliament who helped us to collect all the necessary speech and text data.

7. References

[1] Byrne, W., Hajic, J., Ircing, P., Khudanpur, F., McDonough, J., Peterek, N., and Psutka, J., ”Large vocabulary speech recognition for read and broadcast Czech”, Proc. Workshop on Text Speech and Dialog, Plzen, Czech Republic, 1999, Lecture Notes in Artificial Intelligence, Vol. 1692.

[2] Nouza, J., Nejedlova, D., Zdansky, J., Kolorenc, J., ”Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs”, Proc. ICSLP 2004, Jeju Island, Korea.

[3] van den Heuvel, H., Boudy, J., Bakcsi, Z., Cernocky, J., Galunov, V., Kochanina, J., Majewski, W., Pollak, P., Rusko, M., Sadowski, J., Staroniewicz, P., Tropf, H.S., ”SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed”, Proc. Eurospeech 2001, Aalborg, Denmark.

[4] Žgank, A., Kačič, Z., and Horvat, B., ”Large vocabulary continuous speech recognizer for Slovenian language”, Proc. Text, speech and dialogue : 4th international conference, TSD 2001, Železna Ruda, Czech Republic, Lecture notes in Artificial Intelligence, Vol. 2166, 242–248, Springer 2001.

[5] Dreo, D., ”Slovene speech data base SNABI”, Dialog Man-Machine : second International Workshop, Maribor, Slovenia, 1995.

[6] Žgank, A., Rotovnik, T., Sepesy Maučec, M., Verdonik, D., Kitak, J., Vlaj, D., Hozjan, V., Kačič, Z., Horvat, B., ”Acquisition and annotation of Slovenian Broadcast News database”, Fourth international conference on language resources and evaluation, Lisbon, Portugal, 26th, 27th & 28th May 2004. LREC 2004, Vol. 6, 2103–2106, 2004.

[7] Žgank, A., Verdonik, D., Zögling Markuš, A., Kačič, Z., ”BNSI Slovenian Broadcast News database - speech and text corpus”, Proc. Interspeech 2005, Lisbon, Portugal.

[8] Žgank, A., Verdonik, D., Zögling Markuš, A., Kačič, Z., ”SINOD - Slovenian non-native speech database”, Proc. LREC 2006, Genova, Italy.

[9] Kemp, T., Waibel, A., 2Unsupervised Training Of A Speech Recognizer: Recent Experiments”, Proc. Eurospeech September 5-9, 1999, Budapest, Hungary.

[10] Wessel, F., Ney, H., ”Unsupervised training of acoustic models for large vocabulary continuous speech recognition”. In Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Madonna di Campiglio, Trento, Italy, December 2001

[11] Lamel, L., Gauvain, J., and Adda, G., ”Lightly supervised and unsupervised acoustic model training”, Computer Speech & Language, Volume 16, Issue 1, , January 2002, Pages 115-129.

[12] Nguyen, L., Xiang, B., ”Light supervision in acoustic model training”, Proc. ICASSP 2004. Montreal, Canada.

[13] Gollan, C., Biasni, M., Kanthak, S., Schlüter R., Ney, H., ”Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus”, Proc. ICASSP 2005. Philadelphia.

[14] Biatov, K., Köhler. J., ”Methods and Tools for Speech Data Acquisition exploiting a Database of German Parliamentary Speeches and Transcripts from the Internet”, Proc. LREC 2002, Las Palmas, Spain, June 2002.

[15] Barras, C., Geoffrois, E., Wu, Z. and Liberman, M., ”Transcriber: Development and use of a tool for assisting speech corpora production”, Speech Communication, Vol. 33, Issues 1-2, 5-22, 2001.