The SI TEDx-UM Speech Database: a new Slovenian Spoken Language Resource

Andrej Žgank, Mirjam Sepesy Maučec, Darinka Verdonik
University of Maribor, Faculty of Electrical Engineering and Computer Science,
Institute of Electronics and Telecommunications,
Maribor, Slovenia
E-mail: andrej.zgank@um.si

Abstract

This paper presents a new Slovenian spoken language resource built from TEDx Talks. The speech database contains 242 talks in total duration of 54 hours. The annotation and transcription of acquired spoken material was generated automatically, applying acoustic segmentation and automatic speech recognition. The development and evaluation subset was also manually transcribed using the guidelines specified for the Slovenian GOS corpus. The manual transcriptions were used to evaluate the quality of unsupervised transcriptions. The average word error rate for the SI TEDx-UM evaluation subset was 50.7%, with out of vocabulary rate of 24% and language model perplexity of 390. The unsupervised transcriptions contain 372k tokens, where 32k of them were different.

Keywords: spoken language resource, under-resourced language, unsupervised transcriptions

1. Introduction

Under-resourced languages still present a great challenge for the speech processing community. Slovenian belongs to such a group of under-resourced languages and is, with 2 million speakers, one of the smallest official EU languages. The development of Slovenian speech technology systems started in the end of 80's. Special emphasis was given to the development of spoken language resources. During these years, comparable categories of language resources to the main ones for English were built: SNABI and GOPOLIS, (TIMIT, Resource Management like), 1000 FDB SpeechDat(II) and PoliDat (SpeechDat like), BNSI Broadcast News and SiBN Broadcast News. The common characteristics of these language resources are that they were manually annotated and transcribed. This presents a time consuming and expensive task, especially for under-resourced languages like Slovenian. The first Slovenian language resource that differs from this approach was SloParl with parliamentary debates, which was based on imperfect transcriptions generated in parliament.

The size of these Slovenian language resources is approximately 200 hours of speech, which is significantly less than for frequently spoken world languages. Additional drawback for developing speech recognition systems is the fact that Slovenian language belongs to the group of highly-inflected languages. This group of languages usually needs significantly more spoken training material to successfully develop high quality large vocabulary continuous speech recognition systems.

In order to increase the amount of available language resources for Slovenian automatic speech recognition and other speech technology systems, the decision was made to build a new Slovenian resource based on TEDx Talks (Technology, Entertainment, Design) using automatic acoustic classification and large vocabulary continuous speech recognition. The costs and time usually needed to build such a speech resource were reduced with annotating and transcribing the acquired speech mainly in an unsupervised way.

The paper is organized as follows. The Section 2 presents the acquisition process. The automatic segmentation and transcription are described in Section 3. The manual annotation and transcription of development and evaluation set are presented in Section 4. The finalized speech database is described in Section 5, while the conclusion is given in Section 6.

2. Acquisition

TEDx Talks can be a valuable source of speech material for different categories of languages, especially for under-resourced ones. An example of TEDx Talks speech databases is TED-LIUM corpus (Rousseau et al., 2014). As our goal was to collect as much high quality speech material as possible, we concentrated on talks given by single speakers with clean acoustic backgrounds. The initial source list of talks was collected from the TEDx Slovenian channel on YouTube. The initial source list of talks was manually filtered, where those talks differing from the previously defined talks' characteristics were excluded from the initial source list of more than 300 talks. The more frequent reasons for excluding talks were: frequently overlapping speech, foreign language, music background. The final list of available TEDx Talks given in the Slovenian language contained 242 talks, with a total length of approx. 54 hours of recordings. The selected TEDx Talks were captured from the YouTube in the highest available quality. The captured video was mainly in format H.264. It was used as assistance during the speech database development process. The captured audio was compressed with MPEG AAC codec with different bitrates. The compressed audio signal was converted into WAV audio, as this presents the codec type usually used in other Slovenian speech databases.

The Slovenian TEDx Talks acquired were organized by several organizers within a timespan of more than 6 years. Large number of organizers, were some of them were not backed up by a formal organization, was the main reason

to collect the raw spoken material from YouTube, instead of trying to establish a direct connection with the organizers and content providers as we usually did during our previous language resources development campaigns. This solution significantly eases the initial resource building process, but has a potential disadvantage of not being able to establish a long period connection for continuously acquiring new raw spoken material.

A broad coverage of topics was included within the set, spanning from technology and popular science, social and human awareness, to entertainment. The types of talks were even in the cases where the presenters were scientists, given in a more popular style, as the talks' audiences were usually general. This was the key difference to those scientific talks which are part of a similar Videolectures speech database generated within the scope of the transLectures project (Golik et al., 2013). Another major difference between those two speech databases is that the Videolectures database includes additional text material (slides, conference papers) which can be useful for improving the speech recogniser's language models. Due to the types of talks and the given method of presenting and sharing them, such types of additional material were not available in the case of the SI TEDx-UM speech database.

The comparison from the signal processing point of view between the captured audio signal in the SI TEDx-UM database (AAC lossy codec) and the audio signal in the BNSI Broadcast News speech database (WAV lossless codec), which was applied for training the automatic speech recognition system for producing transcriptions, is given in Figure 1 and 2.

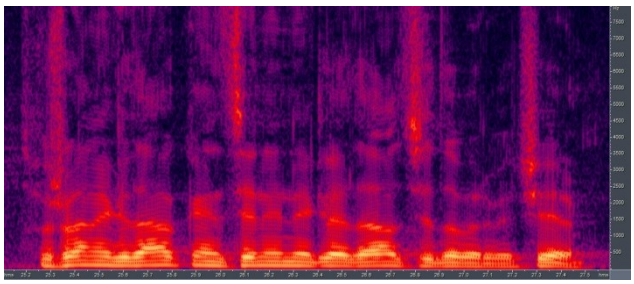

Figure 1: BNSI Broadcast News audio signal.

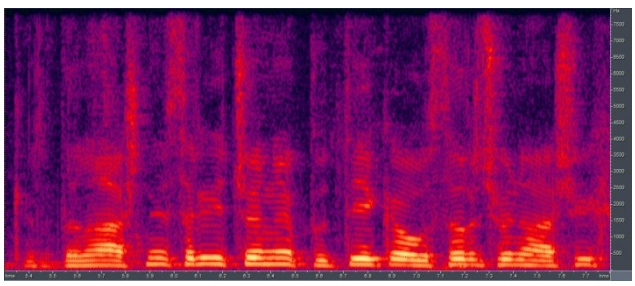

Figure 2: SI TEDx-UM audio signal.

The spectral view shows higher level of noise in case of TEDx recordings as it was recorded in front of audience in larger rooms and halls. Also, some artefacts of lossy speech codec can be observed in case of SI TEDx-UM database.

3. Unsupervised automatic annotation and transcription

The Slovenian large vocabulary continuous speech recognition system was used to produce the unsupervised transcriptions for the training set. This approach enabled us to build the speech resource in a time and cost-efficient way, which is of immense importance for under-resourced languages. The transcriptions done with automatic speech recognition are imperfect and contain different levels of speech recognition errors, the severities of which depend on the topic, speaking style, speaker and acoustic conditions. Nevertheless, several authors (Lamel et al., 2002; Novotney et al., 2009; Rousseau et al., 2014) have already proven that speech resources with automatically generated transcription can be efficiently used in various speech recognition systems.

The automatic speech recognition system (ASR) used for generating transcriptions from captured audio was based on a BNSI Broadcast News speech database (Žgank et al., 2005). The mel-frequency cepstral coefficients and energy features with first and second derivatives were extracted from the audio signal. Additionally, cepstral mean normalization was used to improve robustness. The hidden Markov models (HMM) based on a grapheme acoustic unit were applied for acoustic modeling and the word trigrams for language modeling. The final BNSI Broadcast News acoustic models, used for producing the unsupervised automatic transcriptions were cross-word context dependent grapheme models with mixtures of 32 Gaussian probability density functions per state. More details about the automatic speech recognition system can be found in Sepesy et al. (2013).

The language model was adapted to the broadcast news domain. It was trained on four corpora: BNSI-Speech (transcriptions of BNSI train set), BNSI-Text (collection of different TV scenarios), Vecer (newspaper corpus) and FidaPLUS corpus (reference corpus of the Slovenian written language). Since the corpora were of very different size, each corpus was used to build one component of final model. The final language model was a linear interpolation of all four components, where the interpolation weights were tuned on BNSI development set. The components were trigram models based on Good-Turing discounting and Katz back-off. Singletons were excluded only from the component built on FidaPLUS.

In the first step of generating the automatic transcription for the SI TEDx-UM speech database, the captured audio was segmented into acoustically homogeneous parts, using a GMM based approach. Homogeneous parts were needed for the speech recognition system. The unsupervised transcriptions were produced in the next step, using the above-mentioned final HMM acoustic models.

An example of comparison between the manual transcriptions and automatically generated ones from the SI TEDx-UM evaluation set is given in Table 1.

Transcription quality	Sentence
average	**REF:** lahko vključimo precej simbolike saj veste dobro slabo vroče mrzlo pozitivno negativno **HYP:** ***** pogosto precej simbolike saj veste dobro slabo vročino zelo pozitivno negativno **Eval:** D S S S
low	**REF:** kako se je vrnil pa o tem da smo d danes tik na tem da spet izumre **HYP:** ko se ** ***** ** * vrnejo potem smo * danes tik ** *** zatem spet zmaga **Eval:** S D D D D S S D D D S S

Table 1: Comparison between manual and automatic transcriptions.

4. Manual annotation and transcription

Manual annotation and transcription of speech followed the guidelines specified for the GOS corpus (Verdonik et al., 2013). However, they were reviewed in order to best meet the needs of ASR. The background for the reviewed guidelines is represented in Žgank et al. (2014).

Four levels of speech annotation and transcription are defined: speech metadata, speech segmentation, speech transcription, annotation of acoustic environment and acoustic events. Speech meta-data are provided on various levels: information about the associated audio version, date of transcription, etc.; information about speakers (ID, gender, etc.); information about the communication situation.

Speech segmentation is not a trivial task. Pauses in speech cannot be the only criteria for speech segmentation as they often appear in the middle of syntactically and semantically coherent units. The goal of speech segmentation is that speech segments correspond with so-called utterances in speech, i.e. to semantically, syntactically and prosodically coherent units, while at the same time considering the need of ASR for very precise segmentation. In the case of ambiguity, shorter segments are preferred. The remaining overlapping speech is very carefully separated from non-overlapping speech.

Speech transcription follows the two-level system of orthographic transcription, defined in the GOS corpus. The first level of speech transcription is pronunciation-based, which follows the acoustic forms of words as faithfully as possible; the second level is standardized transcription, which follows the written standard and offers a common form for different pronunciation realizations of the same word form. Table 2 shows an example of the two-level transcription system. Phonetic transcriptions have been added in order to illustrate how was an utterance actually pronounced and how it would be pronounced in standard Slovenian. As a number of data for pronunciation-based and standardized transcription is available via GOS corpus, we used this data to train an automatic procedure for creating standardized transcriptions based on pronunciation-based transcriptions. It has been used to fasten the transcription process for the part of SI-TEDx-UM database that was manually transcribed (development and evaluation subset).

Annotation of acoustic environment in SI-TEDx-UM database is more careful than in the GOS corpus. For ASR it is important that any changes in acoustic environment (background noises, music, speech, etc.) or any acoustic event (breathing, mouth sounds, etc.) are carefully annotated. The Transcriber AG tool was used to make manual transcriptions.

Pron.-b. transcr.	Realized pronun.	Stand. trans.	Standard pronun.	English trans.
mism	m /i: – s m	mislim	m /i: – s l i m	I mean
tko	t k /o:	tako	t a – k /o:	so
kok	k /o: k	koliko	k /o: – l i – k O	how much
s	s	si	s i	did you
rabu	r /a: – b u	rabil	r /a: – b i U	need

Table 2: Example of two-level transcription system with additional phonetic transcription.

5. Speech database

The final speech database covers in total 54 hours of speech. The transcriptions have 372k tokens, 32k of them are different. The SI TEDx-UM speech database is comprised of three subsets: the training subset with unsupervised transcriptions has a total of 51 hours of recordings, while the development and evaluation subsets are needed for the speech recogniser's development and are based on manual annotations and transcriptions. The duration of the first one is 1 hour of recordings and the second one 2 hours of recording.

E&D subset	Topic	PP	OOV
1	travel	409	21%
2	technology	390	23%
3	society	440	22%
4	technology	379	28%
5	art	481	26%
6	society	491	26%
7	science	323	22%
8	science	242	20%
9	society	429	27%
10	art	400	24%
11	society	428	19%
12	science	402	24%
13	society	287	23%
all	various	390	24%
BNSI eval	various	247	4%

Table 3: Perplexity and out-of-vocabulary rate for the SI TEDx-UM and BNSI evaluation set.

The evaluation of automatically generated transcriptions

was done on 3 hours of manually annotated and transcribed evaluation and development subset. The development set was included in the evaluation, as it was not needed in the current speech recogniser setup. The perplexities and OOV rates on different talks are given in Table 3. We can observe considerably higher values in comparison with the in-domain BNSI evaluation subset. It is not a surprise, because language model and vocabulary were adopted to the broadcast news domain. In the future we plan to adapt the language model to the domain of TEDx Talks or at least to use a more general model of Slovenian language. The speech recognition results are given in the form of word error rate (Table 4).

E&D subset	WER(%)
1	50.5
2	54.7
3	57.7
4	39.2
5	67.1
6	46.1
7	52.9
8	35.5
9	51.4
10	35.0
11	52.4
12	70.3
13	38.9
all	50.7
BNSI eval	26.6

Table 4: ASR WER for SI TEDx-UM evaluation and development set.

The average WER for automatically generated transcriptions is 50.7%. This level of speech recognition errors is caused by the difference in domains between the BNSI Broadcast News and SI TEDx-UM speech database, which was already indicated by a high out of vocabulary rate. As comparison, the same speech recognition system achieved 26.6% WER on the in-domain BNSI evaluation subset. The WER variation between different talks is significant. The best one (ID10) achieved 35.0% WER, while the WER for the worst one (ID12) was as high as 70.3%. There is no direct correlation between WER and OOV in case of these two talks. The possible causes for these variations could result from acoustic environment (background, recording channel) or speaker characteristics. As the current speech recognition system works in speaker-independent mode, speaker adaptation procedures could be applied to reduce these variations. The analysis and comparison of acoustic environment will be carried out in future with digital signal processing methods in combination with objective assessment of speech quality.

6. Conclusion

The paper presented a new Slovenian speech resource built in a cost-effective way. This new speech resource will further support the development of speech and language technologies for under-resourced language. The procedures developed during this work will be in future used for building new Slovenian language resources.

The SI TEDx-UM speech resource has the same Creative Common 3.0 license as the TEDx Talks and is freely available to the speech processing research community[1].

7. Acknowledgements

This research work was partially funded by the Slovenian Research Agency ARRS under the contract number P2-0069.

8. Bibliographical References

Golik, P., et al. Development of the RWTH Transcription System for Slovenian. Proc. of Interspeech 2013, Lyon (France), 2013, pp. 3107-3111.

Lamel, L., et al. Lightly Supervised and Unsupervised Acoustic Model Training. Computer Speech & Language 16(1), 2002, pp. 115–129.

Novotney, S., et al. Unsupervised Acoustic and Language Model Training with small Amounts of Labelled Data. Proc. 2009 IEEE Int. Conf. Acoustics, Speech and Signal Process., April 19-24, 2009, pp. 4297-4300.

Rousseau, A., et al. Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), May 2014.

Sepesy Maučec, M., Kačič, Z., Žgank, A.. Speech recognition for interaction with a robot in noisy environment. Przegląd Elektrotechniczny, 2013, 89-5, pp. 162-166.

Verdonik, D., et al. Compilation, transcription and usage of a reference speech corpus : the case of the Slovene corpus GOS. Language resources and evaluation, 47(4), 1031–1048.

Žgank, A., et al. BNSI Slovenian broadcast news database - speech and text corpus. Interspeech 2005, September, 4-8, Lisbon, Portugal.

Žgank, A., Zwitter Vitez, A., Verdonik, D. The Slovene BNSI broadcast news database and reference speech corpus GOS: Towards the uniform guidelines for future work. Ninth International Conference on Language Resources and Evaluation, May 26-31, 2014, Reykjavik, Iceland. pp. 2644-2647.

9. Language Resource References

University of Maribor (2006). Slovenian BNSI Broadcast News Speech Corpus. Distributed via ELRA, 1.0, ISLRN 502-280-144-938-4.

University of Maribor (2016). Slovenian SI TEDx-UM Speech Database. freely available, 1.0, ISLRN requested.

[1] SI TEDx-UM homepage: http://ietk.feri.um.si/en/portfolio/sitedxumenglish