BulPhonC: Bulgarian Speech Corpus for the Development of ASR Technology

Neli Hateva, Petar Mitankin, Stoyan Mihov
FMI Sofia University, FMI Sofia University and IICT Bulgarian Academy of Sciences, IICT Bulgarian Academy of Sciences
FMI 5 James Bourchier Blvd. 1164 Sofia Bulgaria, IICT Acad. G. Bonchev St. Block 25A 1113 Sofia Bulgaria
nelly.hateva@gmail.com, petar@lml.bas.bg,stoyan@lml.bas.bg

Abstract

In this paper we introduce a Bulgarian speech database, which was created for the purpose of ASR technology development. The paper describes the design and the content of the speech database. We present also an empirical evaluation of the performance of a LVCSR system for Bulgarian trained on the BulPhonC data. The resource is available free for scientific usage.

Keywords: Language resources, speech recognition corpus, Bulgarian speech database

1. Introduction

A crucial language resource for building ASR systems is a properly created speech corpus, which is phonetically annotated and representative. Recently a number of speech corpora for less common languages for ASR development have been created (Odriozola et al., 2014; Pinnis et al., 2014; Korvas et al., 2014). For Bulgarian ASR only few speech resources have been created – (Tanja Schultz, 2002; Mitankin et al., 2009). Both of the resources have been recorded in non soundproof environment. The number of speakers and number of utterances in those databases have been relatively small compared to ASR speech databases for other languages[1].

Here we present a new speech corpus created in the framework of the AComIn project[2]. The corpus has been specially designed to facilitate the development of modern ASR technology. In the next section we provide a detailed description of the database. Afterwards an evaluation of the performance of the ASR system (Mitankin et al., 2009) trained on the BulPhonC data is given. Finally we present availability information and the conclusion.

2. BulPhonC: Bulgarian Speech Database

2.1. Phonetic Balancedness of the Speech Recognition Corpus

The phonetic annotation of the BulPhonC corpus has used the phonetic system presented in (Mitankin et al., 2009), given on Table 1.

The corpus consists of speech read from selected declarative and interrogative sentences. Table2 shows some examples of sentences in the corpus. The sentences in the BulPhonC corpus were automatically selected from a set of 21838 sentences, which were phonetised with the Speech-Lab text-to-speech system for Bulgarian, (Andreeva et al., 2005).

На първото стъпало са разположени фундаменталните физиологически нужди - хранене и възпроизвеждане.
Спомням си, че когато веднъж в Белград исках да си купя часовник, продавачът ме запита колко часовника.
Кой футболист държи рекорда за най-много отбелязани голове в рамките на едно световно първенство?
След службата в църквата свещеникът хвърля кръст във вода, а ергени го изваждат.

Table 2: Example of sentences in the BulPhonC corpus

First, from the whole set of 21838 sentences we automatically selected 422 which contain at least three times almost each phoneme bigram possible for Bulgarian language. In particular, in the 422 sentences from all 806 possible bigrams only 58 have less than 3 occurrences. These 58 bigrams are very rare in Bulgarian, for example “c4”, “69”, “4v”. Afterwards we manually reduced the 422 sentences to 319 regarding the representativeness in order to shorten the time needed for a reader to read them all. The average number of words in a sentence is 10.74. The size of the sentences is comparable to the size of the text in the BAS PhonDat PD1 corpus, (Draxler, 1995).

2.2. Distribution of Speakers

The total number of the recorded individuals in the BulPhonC Bulgarian speech database is 147. For each of them we collected information about their gender, age, higher education (if any) and location in which they completed their secondary education.

The representativeness of the speakers’ age and regional distribution in the corpus is limited because of the shortage of available volunteers.

Distribution of Speakers with Regard to Age and Gender

From the recorded people 85 are females and the rest 62 are males. BulPhonC contains speech from speakers of different ages between 11 and 62. However, the speakers below 18 years are only 3: one aged 11 and two aged 16. The distribution of the speakers with regard to age is given in

[1]The number of speakers in the corpus in (Tanja Schultz, 2002) is 77 and each speaker read 1940 words. The number of speakers in the corpus in (Mitankin et al., 2009) is 46 and each speaker read 478 words.

[2]AComIn “Advanced Computing for Innovation”, grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions)

A	мерак	l	лек, лампа	g	град
a	мазе, кедър	m	мама	d	дар
e	тел, перо	n	нар	v	жар
i	бик, пирон	p	пек	z	зар
O	конче	r	ръка	k	кана
o	борба, кичур	s	син	4	чар
U	Тунис	t	тих	6	шах
Y	катър	f	фар	j	край, Кольо
b	баба	h	хол	0	чорбаджия
w	Варна	c	цар	9	годзила

Table 1: BulPhonC phonemes

	Female	Male	Total	%
11-23	17	14	31	21
24-54	48	37	85	58
55-62	20	11	31	21
Total	85	62	147	100
%	57.82	42.18	100	

Table 3: The distribution of the speakers with regard to age and gender

	Speakers	%
Sofia	87	59.18
Plovdiv	10	6.80
Varna	6	4.08
Stara Zagora	5	3.40
Vratsa	5	3.40
Burgas	4	2.72
Dobrich	4	2.72
Silistra	4	2.72
Veliko Tarnovo	4	2.72
Others	18	12.24

Table 4: The distribution of the speakers with regard to the location in which they completed their secondary education

Table 3.

Distribution of Speakers with Regard to Region
All of the recorded people except one are native Bulgarian speakers. The distribution with regard to the location in which they completed their secondary education is given in Table 4.

2.3. Recording Platform

The recordings have been done in a soundproof room. All of the speakers were recorded with three condenser microphones:

- unidirectional Sennheiser MK4;
- omnidirectional Behringer ECM8000;
- bidirectional SM Pro Audio MC03.

The signals were recorded at 48KHz, 24bits on a TASCAM DP32 digital recorder. Additionally 90% of speakers were recorded with a electroglottograph Glottal Enterprises EG2-PCX and 48% - with a digital stereo camera DXG DVX5F9.

2.4. Validating and Cleaning the Utterances

We applied a semiautomatic approach for validating and cleaning the recorded long audio wave files containing many spoken utterances. First, we automatically segmented the long wave files into utterances and annotated each utterance with its corresponding sentence. Then we manually verified the result of the automatic segmentation and threw out all erroneously annotated utterances. It turned out that 11.21% of the utterances were not segmented or annotated correctly.
The automatic procedure for segmentation and annotation of utterances makes use of the SpeechLab text-to-speech system for Bulgarian, (Andreeva et al., 2005), and the speech recognition (SR) system for Bulgarian presented in (Mitankin et al., 2009). SpeechLab generates a set of possible phonetisations for each sentence.
For each sentence we build a specific HMM model, which we use to find in the long wave file the utterance that corresponds to the sentence. The output of this process is a sequence of words from the given sentence and their corresponding time boundaries in the signal. We assume that the required utterance in the signal will be recognised with high accuracy and as a result in the output sequence of words we will have a subsequence which is close to the given sentence. With the extracted utterances we adapt the SI acoustic model to the given speaker and then we apply again the above procedure for all sentences with the adapted acoustic model. The automatically obtained segmentation is then finally verified manually.

2.5. Speech Corpus Annotation

Each manually verified utterance is automatically annotated on phoneme level. The annotations are kept in a format supported by praat, (Boersma and Weenink, 2016). Fig-

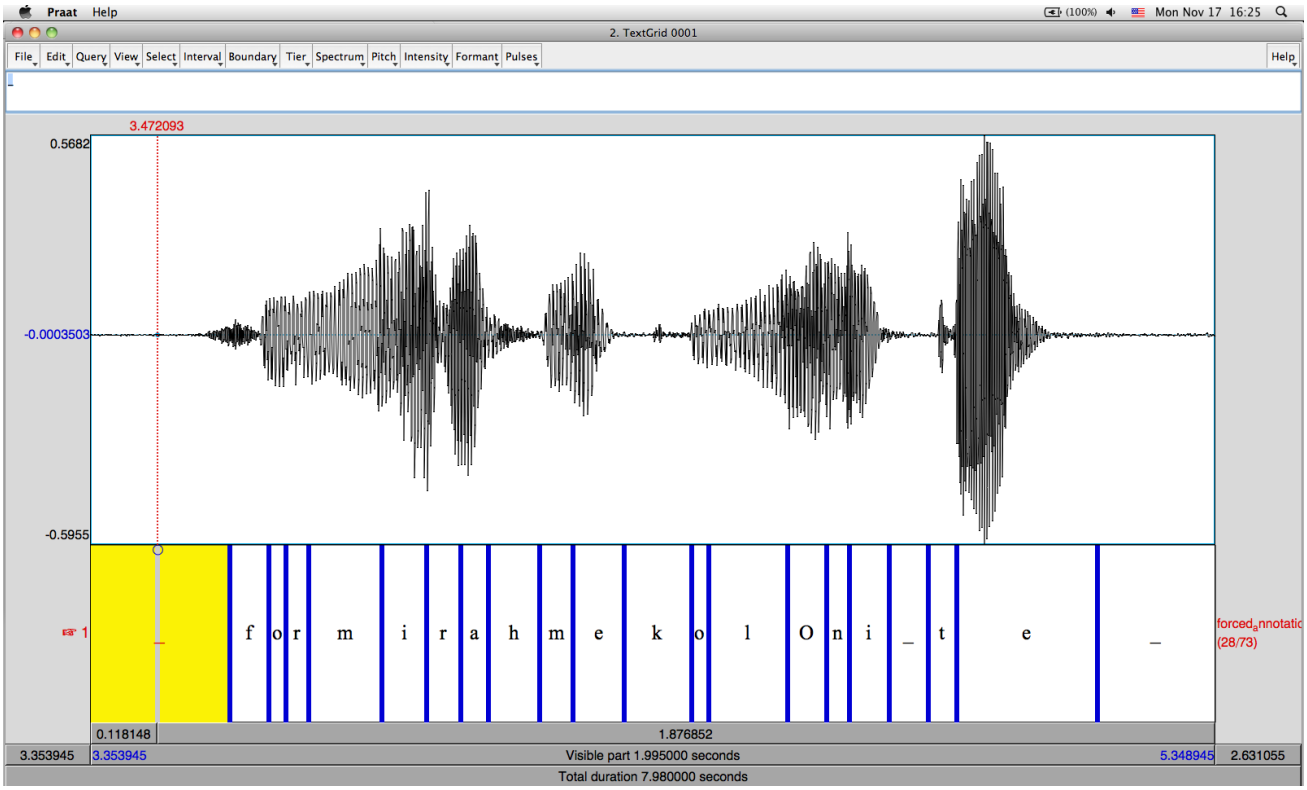

Figure 1: Utterance annotation in Praat.

ure 1 shows an utterance and its corresponding phoneme annotation.

2.6. Speech Corpus Statistics

The speech corpus consists exclusively of read speech, 319 sentences divided into two parts. The first part contains 148 and the second one the remaining 171 sentences. All of the 147 speakers read the first part, 23 of them also read the second part. So, the total number of the recordings is 170. Each one lasts between 15 and 30 minutes at normal pace of reading. The total number of the pronounced sentences is 24801. 2762 sentences were dropped after cleaning the data. The remaining 21891 (88.79%) sentences are included in the corpus. Their total length is around 32 hours after cleaning the pauses between the utterances. The final structure of the corpus were taken from (Draxler, 1995), namely: signals recorded in the pronunciation of individual sentences are saved in separate files along with their phonetic annotations. For each reader the files are kept in a separate directory. The speech corpus is the aggregate of these directories.

3. Evaluation

We have tested the new method using our LVCSR system for Bulgarian (Mitankin et al., 2009). All tests are performed using a speaker independent (SI) acoustic model trained on the BulPhonC corpus. The n-gram language models were constructed using a ∼250M words legal corpus for Bulgarian, created in the framework of the project for development of Bulgarian ASR system for juridical texts (Mitankin et al., 2009). The test set consists of 9 speakers with 50 long legal utterances each. We have varied the beam width for the beam search between 1000 and 3000 states by a step of 500. For building the lattice a bigram language model was used. The rescoring has been performed using a trigram language model.

	Word Accuracy	Time ratio
Beam=1000	92.63%	0.40x
Beam=1500	93.57%	0.54x
Beam=2000	93.78%	0.74x
Beam=2500	93.82%	1.02x
Beam=3000	93.85%	1.37x

Table 5: Performance of the Bulgarian Speech Recognition System.

Table 5 presents the resulting word accuracy (1 – WER) and time ratio for the recognition. The measurements are performed on the recognition by varying the beam width.

4. Availability

The BulPhonC corpus is available in compressed tar.gz format. The archive contains 16-bit PCM wave files with 16 kHz sampling frequency (one file for each utterance), the text of the utterance and their corresponding phoneme annotation files in Praat TextGrid format. The 16 kHz wave files were obtained by downsampling the original 48 kHz recording files.

The corpus is available free for scientific usage. Requests for obtainment have to be sent to

BulPhonC@lml.bas.bg. More information can be found on http://lml.bas.bg/BulPhonC.

5. Conclusion

In this paper we have presented the Bulgarian speech corpus BulPhonC created for the development of ASR technology for Bulgarian. The quantitative and qualitative characteristics of the database have been shown in detail. Furthermore, the results of speech recognition experiments on the BulPhonC database have been presented, in order to evaluate the applicability of the database for the development of ASR technology.

6. Acknowledgments

The research leading to these results has received funding from the research project AComIn "Advanced Computing for Innovation", grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions). We also thank the anonymous reviewers for their helpful comments and suggestions.

7. Bibliographical References

Andreeva, M., Marinov, I., and Mihov, S. (2005). Speechlab 2.0: A high-quality text-to-speech system for bulgarian. In *Proceedings of the RANLP International Conference 2005*, pages 52–58, September.

Boersma, P. and Weenink, D. (2016). Praat: doing phonetics by computer [computer program]. version 6.0.14, retrieved 11 february 2016 from http://www.praat.org.

Draxler, C. (1995). Introduction to the verbmobil-phondat database of spoken german. In *Prolog Applications Conference PAP 95, Paris*. http://www.bas.uni-muenchen.de/forschung/Bas/BasPD1eng.html.

Korvas, M., Plátek, O., Dušek, O., Žilka, L., and Jurčíček, F. (2014). Free english and czech telephone speech corpus shared under the cc-by-sa 3.0 license. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Mitankin, P., Mihov, S., and Tinchev, T. (2009). Large vocabulary continuous speech recognition for bulgarian. In *Proceedings of the RANLP 2009*, pages 246–250, September.

Odriozola, I., Hernaez, I., Torres, M. I., Rodriguez-Fuentes, L. J., Penagarikano, M., and Navas, E. (2014). Basque speecon-like and basque speechdat mdb-600: Speech databases for the development of asr technology for basque. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Pinnis, M., Auziņa, I., and Goba, K. (2014). Designing the latvian speech recognition corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Tanja Schultz. (2002). Globalphone: A multilingual speech and text database developed at karlsruhe university. In *Proceedings of the International Conference of Spoken Language Processing, ICSLP 2002*, September. http://islrn.org/resources/250-105-856-478-2.