Building a Speech Repository for a Serbian LVCSR System

Siniša Suzić, Stevan Ostrogonac, Edvin Pakoci, and Milana Bojanić

***Abstract*—This paper describes the procedure of collecting speech and corresponding textual data and the processing needed to create a repository for training a LVCSR system for the Serbian language. The speech database for Serbian consists of speech recordings from audio books, radio programmes and talk shows, as well as read utterances from an array of male and female speakers. Currently, approximately 200 hours of speech recordings are collected, together with corresponding orthographic transcriptions which contain around 200 thousand words (over 3 million characters).Audio files are split in order for each of them to contain a single utterance. The corresponding transcriptions are used to create label files as well as for training the language model (LM) – namely, new transcriptions are added to the existing textual corpus earlier collected for the purpose of creating the LM. The software which was specially designed for building the speech repository for Serbian is also briefly described.**

***Keywords*—large vocabulary continuous speech recognition, Serbian, speech database.**

I. Introduction

Large vocabulary continuous speech recognition (LVCSR) is a basis for various speech technologies applications. In order to build an efficient LVCSR system, high-accuracy acoustic models, a large-scale language model and an efficient decoder are essential. But, in contrast to many other new technologies, these resources, except for the decoder, must be developed for each language separately.

The work presented in this paper represents an extension of the ongoing research on the development of a LVCSR system for Serbian at the Faculty of Technical Sciences in Novi Sad in collaboration with AlfaNum Ltd, Novi Sad. One implementation of the LVCSR decoder has already been presented [1], as well as the language model which has been created for Serbian[2].

In general, whenever a new system is designed, representative speech data needs to be collected and manually transcribed ,which is a time-consuming and costly process. Commonly, an annotated speech corpus and a pronunciation dictionary are used for acoustic model training. The accuracy of the system depends mostly on the amount and quality of the training data. In order to obtain robust acoustic models, their parameters have to be estimated on very large corpora, involving many speakers selected in a way which represents a typical distribution of age, gender and dialect so that the corpora would include as much speech variability present in a particular language as possible. More details about the data requirements for automatic speech recognition can be found in [3].

Prior to this research, the only database for automatic speech recognition (ASR) for Serbian, SpeechDat(E), consisted of telephone channel recordings [4]. Some of the existing LVCSR speech corpora for other languages are presented in [5], [6] and [7]. In this paper, the applied procedure of data collection and processing necessary in order to obtain appropriate corpora for training acoustic models is described in detail along with the database structure.

The rest of the paper is organized as follows. Section 2 gives a description of sources from which the data was collected and the process classifying the materials. Section 3 describes the process of database preparation in detail together with the software tools developed for the purpose of creating speech repositories. In Section 4, overall content of the database is presented, while in Section 5 the plans for future database improvements are discussed.

II. Data Acquisition

Creating a large speech database can be a great logistic task. Firstly, appropriate recording equipment and a studio need to be provided. Secondly, many different speakers need to be included in the recording process. In order to reduce costs, a decision has been made not to record all the data but to use some already existing audio material.

The first and the most valuable source of audio material was the Audio Library for the Blind “Dr Milan Budimir” in Belgrade. This institution possesses a large collection of audio books. The books are read by professional speakers in studio environment. The majority of these audio books are available in *.mp3* file format. Besides audio books

Paper received March 12, 2014; revised July 8, 2014; accepted August 21, 2014. Date of publication November 15, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Branimir Reljin.

This paper is a revised and expanded version of the paper presented at the 21th Telecommunications Forum TELFOR 2013.

This work was supported by the Ministry of Education, Science and Technological Development of Serbia within the Project "Development of Dialogue Systems in Serbian and other South Slavic Languages" (TR-32035).

Siniša Suzić, AlfaNum – Speech Technologies, Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia (phone 381-21-475-0204, e-mail: sinisa.suzic@alfanum.co.rs).

Stevan Ostrogonac,Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia (phone 381-21-475-0204, e-mail: ostrogonac.stevan@uns.ac.rs).

Edvin Pakoci, AlfaNum – Speech Technologies, Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia (phone 381-21-475-0204, e-mail: edvin,pakoci@alfanum.co.rs).

Milana Bojanić, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia (phone 381-21-475-0204, e-mail: bojanic.milana@yahoo.com).

from the library “Dr Milan Budimir”, other freely available audio books were also used. This data consists of high quality recorded speech – with a minimal amount of background noises and with a very high percent of correctly pronounced words and phones. The sole quantity of material, which adds up to around 190 hours, is enough to render this database segment as extremely important to this speech repository for Serbian.

The second type of material which was collected are recordings of radio shows which are publicly available on internet sites of radio stations such as “Radio Belgrade”, “B92”, etc. This material includes recordings of news and talk shows, which are also available in *.mp3* file format. This segment of the database is significant because it contains spontaneous speech such as dialogues between the host and the guests in radio talk shows and similar programmes. Joining this radio database with the rest of the material can greatly improve the quality of continuous speech recognition systems designed for or applied to spontaneous, unconstrained speech segments.

Lastly, a significant amount of data consists of recordings of utterances read by lots of different male and female speakers. The set of utterances was constructed for speech recognition purposes in a way to include all possible phones and common phone combinations present in Serbian in a relevant number of appearances. Speakers were instructed to read the utterances very clearly and to articulate all given words and phones correctly. These utterances consist of numbers, names, sequences of individual words and short sentences. This database segment was intended to be used in acoustic model training for continuous speech recognizers developed earlier within our research group. Therefore, it has already been recorded.

Before any of the described materials were included in the database, they had to be manually reviewed. If any kind of damage in certain recordings was detected, such as prominent noises or any kind of equipment malfunction, those recordings were not processed further. Parts of the radio database which had more than a single speaker talking at a particular time were excluded as well. In addition, attention was paid to the recording parameters – sampling frequency and number of bits per sample, as well as the effective spectral content which will be explained in the following section.

III. Preparation of the Speech Repository

The process of database preparation applied to the last group of audio recordings differed from the preparation of the rest of the database, since that group was already processed in an appropriate manner prior to the work focused on this new database. All these differences will be mentioned as well.

The preparation of the speech recognition database comprises several steps, which are presented in Fig. 1. After an audio file is manually reviewed and labeled as not damaged, it is divided into smaller segments. Each of these parts is now a separate wave file containing a single spoken sentence, and it is saved in *.wav* file format. All obtained sentence files are associated with a corresponding description *.txt* file, which holds information about the original audio file: sampling rate, bits per sample, channel quality, gender of the speaker, type of audio source (audio book, radio news, radio talk show), audio format (e.g. PCM),etc. Additional properties, that include data such as effective spectral content (which usually depends on the recording type, as audio book recordings have generally a wider spectrum, while e.g. the frequency range of radio database recordings is generally narrower), name of the speaker, name of the person who split the original file, name of the person who carried out the revision of the selected database segment, and so on, are also included. If any of the mentioned properties are unknown, the appropriate field is left empty. Information from these descriptive files can be used in the training phase of the ASR system to include only audio files with specific properties.

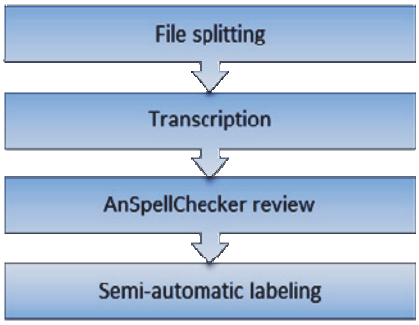

Fig.1. Database preparation process.

Therefore, the first step of the database preparation is splitting audio files into smaller ones which represent single sentences. The simplicity of this procedure mostly depends on the type of audio material. In the case of audio books the procedure is relatively easy. Sentences are correctly pronounced and the boundaries between them are clearly noticeable. The situation is similar with radio news recordings, which are also read by professional speakers. On the other hand, the work on splitting files is much more complicated in case of radio talk shows which contain spontaneous speech. In many situations it is very hard to detect and separate the end of one and the beginning of another sentence. In addition, there is a lot of speaker overlap, so these sections of data have to be either cut out (if they last for any significant amount of time) or marked as flawed (if this refers to only a word or just a few successive words). As it was mentioned before, the part of database consisting of uttered words and short sentences by different speakers was already processed – all utterances already existed as separate files.

The next step in the database preparation process is the transcription of spoken words, which is done manually. In the Serbian language, words are read the way they are written. The main rule in this step is that the phonetic transcriptions of all the words should be noted as heard by human listeners (not as provided by dictionaries, but as the particular speaker has pronounced them in the given sentence). The correct surface forms of words needed for extracting textual content from the transcriptions are also provided in a way which will be explained shortly.

Furthermore, several additional conventions and rules had to be established for transcribing certain word categories, so there would not be any disagreements between different transcribers. These conventions also enable applications to successfully check formed transcriptions and create appropriate labels for each audio file. These word categories include:

1) *numbers* – they are not written using digits. Instead, their phonetic transcription is given, so they could be broken into appropriate phonemes afterwards;

2) *abbreviations* – they are written using capital letters. If the abbreviation is pronounced as a word it is written as such (e.g. “DOS” or “MUP”, like any other word). In the opposite case, when it is spelled out ,it is written letter by letter as well, while the origin (word form) of the abbreviation should follow in brackets (e.g. “D O S (DOS)” or “M U P (MUP)”). This will be explained in more detail in the rest of this section;

3) *foreign words or native words spoken in linguistically unorthodox manner* – they are written as they are spelled in Serbian within square brackets. If these words are frequently used in this form in Serbian and are already incorporated in the AlfaNum morphologic dictionary [8], the square brackets do not need to be used. If a word (or a sequence of words) is marked with square brackets, the transcription of that word (or a word sequence) must be followed by the original transcription (or what is considered to be the semantically and grammatically correct transcription) within regular brackets. The motivation for marking words with square brackets will also be explained in the rest of this section.

For ASR purposes, only some of the basic punctuation marks are used: full stops, question marks, exclamation marks and commas. Some special marks (tags) are also used. They are listed below.

1) <*spk*>This tag is used for speaker-related noises (e.g. if the speaker stuttered during speech or uttered something incomprehensible);

2) <*int*>This tag is reserved for noises caused by a source other than human, perceived between speech segments (e.g. telephone ringing, opening of a door, some buzzing caused by the recording equipment, etc);

3) <*some_word*>This tag marks words which are damaged in the acoustic sense. The damage can be caused by some external factors(e.g. music, or another person who started to talk in the background),or by the speaker himself(e.g. he or she did not articulate all the phones, or did not pronounce some of them correctly). By convention, these words are written the way they sound, which means that the sequences of letters may not represent words contained in the dictionary. Afterwards, when these words are broken into phonemes, those phonemes will be marked as damaged and used only as context for adjacent phonemes. A word marked with this tag is always followed by its correct (dictionary) form within regular brackets;

4) *[some_word]* Square brackets mark either foreign words or mispronounced words if all the phones are articulated correctly. Words marked using this tag are also always followed by their correct form within regular brackets;

5) *( )*This tag is used after the tags [ ] and <>. The word (or a sequence of words) within these brackets is the authentic, dictionary form of the damaged or mispronounced word which preceded it. This information is important because it allows the transcription to be used for training language models for the ASR system.

In a special case, when the word inside [ ] or < > tags is actually pronounced as given in the dictionary, a shorter way of writing it down is allowed – instead of having to write the word twice, first inside square or angled brackets, and then again inside regular brackets, the transcriber can write the word inside a double tag, such as <( )> or [( )]. This simply shortens the written form of an utterance.

One example of a transcribed file is given in Fig. 2. It can be seen that different combinations of tags are also allowed in cases when the transcription within < > or [ ] tags matches the dictionary form. This makes the work easier for people who are writing transcriptions.

<spk> A koliko su ovi <(društveni)> sajtovi <spk> štetni, odnosno korisni, za decu?

Fig. 2. An example of a transcription of an audio file.

As with file splitting, the work on transcribing files is much more complicated in case of radio talk shows than in case of audio books, radio news and chosen uttered words and short sentences. In talk show recordings, the percentage of mistakenly pronounced and damaged words is much higher.

When the transcription is finished, an additional check is carried out using a software tool called *AnSpellChecker* [9]. This software tool detects spelling errors for Serbian. It is based on the Serbian morphological dictionary, which contains a number of foreign words as well, so if those words are found in the dictionary, they are not marked as erroneous. The *AnSpellChecker* application ignores the text within < > and [ ] tags as well as words written in capital letters (words spelled letter by letter or abbreviations). This spellcheck has proven to be very useful even though it does not eliminate all typing errors. It ensures that the transcriptions are accurate enough to be further used in the training of language models. Correct transcriptions are of course also very important in the process of acoustic model training because they prevent wrong data to be used for training a specific phoneme model. However, *AnSpellChecker* cannot detect grammatical errors. For example, if a noun is written in a wrong case and it can be found in the dictionary in the given form, this software tool will not recognize it as erroneous.

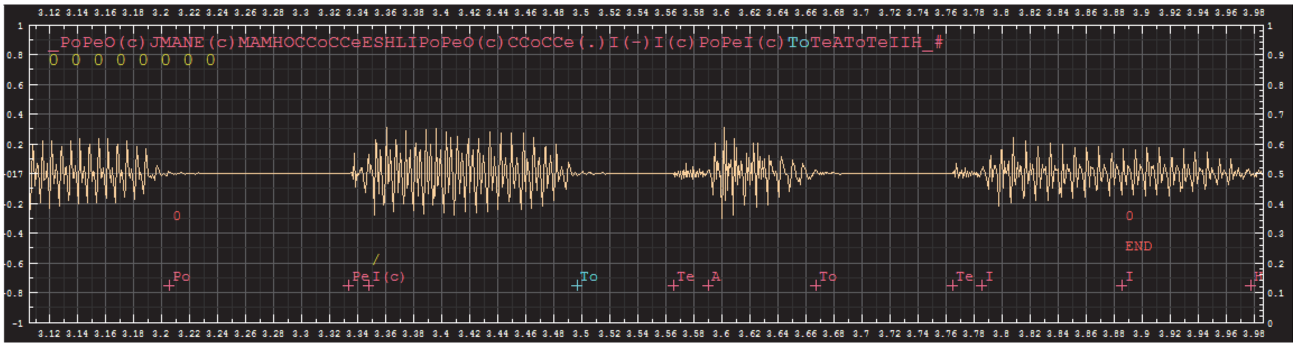

Fig. 3. The interface of the *SpeechLabel* application.

The final step in database preparation is semi-automatic labeling. For this process, a tool called *AnLabelCreator* has been developed. As input, this tool takes the input transcriptions in UTF-8 format and corresponding descriptive files with wave file properties. The application first processes the transcriptions file – excludes the wave file names and words inside *( )* tags, as well as <*spk*> and <*int*> tags and all brackets (while saving at what position in utterance they occur). Then, for each input wave file, *AnLabelCreator* creates appropriate labels using the following procedure:

1) A developed part-of-speech (POS) tagger [10]determines an array of properties for each word in the utterance, including its POS, base form (lemma), values of particular morphological categories depending on the POS, as well as accentuation;

2) For each word, functions developed within a dynamic-link library (DLL) are used to generate possible pronunciations for it in Serbian. Among the given pronunciations, it chooses the one that matches the POS tagger output in the best way. The chosen pronunciation is then sent to the transcriptor module, which modifies it according to an input file containing transcription rules. These are all specialized for Serbian and for the given application. One kind of rule is e.g. breaking of plosives (stops) and affricates into occlusion and explosion parts – this approach was, however, abandoned for ASR purposes. Other rules include substitutions of different sequences of phonemes with other sequences according to how people actually pronounce certain phones in the given order. The given set of rules was developed for Serbian during several years of experience with the development of ASR and text-to-speech systems;

3) The final pronunciations of each word are transformed into appropriate labels. The wave file name is written relative to the desired speech database directory. Each phoneme is then written with appropriate attributes, such as *stressed* (according to accentuation given by the POS tagger) or *damaged* (according to < > tags in the transcription). Also, starting and ending silences are added, as well as <*spk*>and <*int*>labels at appropriate positions. For word-starting phonemes, the POS tagger output for that word is appended as a special property field. All phoneme starting times and quality metrics are initialized to zero values.

After all the files are processed, boundaries between individual phones are calculated using forced alignment. Starting from a small portion of the database previously manually labeled and previously trained acoustic models used for alignment (trained using the manually labeled part of database), phoneme segmentation for the whole database is determined automatically in this manner. Alignment acoustic score for each phoneme is saved in the phone quality field. Silences are afterwards additionally corrected based on the signal energy.

Finally, the generated label files are corrected manually using the AlfaNum *SpeechLabel* software [11]. Fig. 3 depicts a part of one labeled file in *SpeechLabel* software. Each label file can be checked by simultaneous listening to the uttered sentence and looking at its transcription (or specific parts of it). This application allows the user to quickly find possibly problematic parts of the automatically labeled files using the search by particular phonemes and acoustic scores assigned to them during the realignment process. Naturally, the lower the score, the more it is likely that the phones in that part of the file are not aligned well. Besides changing phone boundaries manually, the application also enables tagging phonemes as damaged (or un-tagging them as well), inserting or deleting phonemes where necessary and assigning an array of attributes to certain phonemes.

For the purpose of extracting textual content from the transcriptions for language modeling, an application called *LMCorpusExtractor* has been developed. This application eliminates the text within < > and [ ] tags, as well as all the tags from transcriptions. It also erases the names of the audio files from the beginning of each transcription. Furthermore, special cases are also treated (double tags, tagged sequence of words where a subsequence is marked with another tag, etc). Punctuation marks immediately following any of the tags sometimes need to be erased together with the tagged text, while sometimes need to be joined with the previous word (e.g. punctuation at the end of utterances). If any of the tags are inserted incorrectly, the transcription of the corresponding audio file is discarded from the language model textual corpus. It should be noted that the described process is slow, and therefore not optimal for creating textual corpora for language model training, but it is useful because the textual content obtained in this way is "clean" in the sense that it does not contain a wide range of special characters, which is the case with the content that may be obtained by copying textual data from books, web sites or other sources.

IV. Database Content

The current database consists of approximately 190 hours of audio book recordings and more than 6 hours of talk show recordings, with almost another 4 hours of radio news. The uttered words and sentences part of the database adds up to around 14 hours of material.

Audio book material comprises more than 90000 sentences, and is approximately equally distributed between male and female speakers. There are 25 male and 35 female speakers for which the names are known. Beside these, there are several unidentified speakers. There are 7 audio books read by unidentified male speakers, and 9 audio books read by unidentified female speakers. It may be that some of those books are read by the same speaker (or by one of the identified ones), but this is considered unlikely. This adds up to almost 80 speakers in this part of the database, with an average of about two and a half hours of recordings for each of them. The difference in the number of male and female speakers is compensated by taking into count more data from each male speaker than from a female speaker. This results in a fair balance between genders within the audio database.

As for the radio database, more than 3000 utterance audio files were created. Around 1300 of them are from radio news, and the rest are from talk shows and similar radio programmes. The difference between this kind of material and the audio books is a lower sampling rate (11 kHz compared to 22 kHz and 44 kHz) as well as poorer spectral content (little or no frequency content above 5 kHz due to different format conversions) in most of the material, which may lead to the need for exclusion of this part of the database from acoustic model training and the exclusive use of it in specific training of spontaneous speech models. All in all, around 10 hours of radio material is present. Speakers are known for the talk show part of the base – in total there are 9 male and 6 female speakers. In the news segment, there is also a mix of speakers of both genders.

As mentioned before, the final part of the database consists of read sequences of words and short sentences from lots of different speakers. A smaller part of this database segment consists of just six fixed long word sequences for 4 male and 7 female speakers. The rest was made by having the speakers say the following types of utterances – their name and identification number, two sequences of numbers, 10 sequences of around 5 individual words each and 70 short sentences. A total of 121 male and female speakers (about the equal number of speakers for each gender) took part in recording this database segment. In total, this segment added extra 14 hours of studio-quality read material to the database. Transcriptions related to this part of the database are not intended to be used in language model training, since all speakers recorded the same chosen phrases.

The overall content of the database is shown in Table 1 and Table 2. Letters 'A', 'B' and 'C' used as column labels denote different parts of the database – 'A' corresponds to audio books, 'B' is related to the radio segment, while 'C' is the last, pre-recorded part.

Table 1: Number of speakers: identified + unidentified.

	A	B	C
Male	25+7	9	60
Female	35+9	6	61

Table 2: Amount of audio material (in hours).

	A	B	C
Male	100	6	7
Female	90	4	7

In Table 3, the phonetic content of the database is presented. The phonemes are given in groups determined by the manner of articulation – vowels, affricates, fricatives, stops, nasals, laterals, vibrants and approximants (semi-vowels). Each phoneme and group is linked to a value representing its percentage among all phonemes in the database (transcriptions), as well as a value derived from a large textual database for Serbian used for language modeling. The latter value is given for comparison. This table shows that the database is well balanced phonetically. The small statistical differences have probably arisen from the fact that the LM corpus includes a wide variety of functional styles, while in this speech database the narrative style dominates all others, as its biggest part is comprised of audio books.

V. Conclusion and Further Research

A good speech repository is of great importance for developing a high-quality LVCSR system. The repository described in this paper can be used for training good acoustic models. Our future plans are to extend the described database which will definitely contribute to improving the performance of the ASR system for Serbian. One way of achieving this goal is by using an automatic transcription system [12]. The downside of this approach could be that automatic annotation commonly causes more transcription errors than manual transcription. Of course, additional checking of the phone boundaries may improve the quality of the resulting acoustic models, but it has been decided to rely on having the labels realigned better by using an iterative procedure of model training – after the initial models are trained, the database is realigned using those models, which is followed by another round of training, and so on. Since the collected data differs in quality, it should not be all mixed together when certain acoustic models are being trained. The deficient type of data in the current database are recordings of telephone channel signals. Our current efforts are focused on obtaining more data of this specific type and quality in order to enable the development of good ASR systems which are going to be used on mobile phones.

TABLE 3: DATABASE (DB) PHONETIC CONTENT ANDCOMPARISON WITH THE LM CORPUS.

PHONETIC CATEGORY	REPRESENTATION [%]		PHONEME	REPRESENTATION [%]		PHONEME	REPRESENTATION [%]	
	DB	LM		DB	LM		DB	LM
vowels	45.14	44.33	A	12.39	11.99	O	9.78	9.09
			E	9.52	9.68	U	4.33	4.23
			I	9.12	9.34			
stops	18.05	18.65	P	2.85	2.98	B	1.57	1.54
			K	3.5	3.46	G	1.69	1.59
			T	4.48	4.76	D	3.96	4.32
affricates	2.79	2.77	C	0.73	0.87	DŽ	0.07	0.04
			Č	1.1	0.92	Đ	0.24	0.25
			Ć	0.65	0.69			
nasals	9.78	9.5	M	3.68	3.09	NJ	0.7	0.68
			N	5.4	5.73			
vibrants	4.53	5.36	R	4.53	5.36			
fricatives	9.24	9.41	F	0.19	0.28	H	0.64	0.66
			S	4.94	5.23	Z	1.7	1.78
			Š	1.22	0.98	Ž	0.57	0.48
laterals	3.74	3.2	L	3.19	2.7	LJ	0.55	0.5
approximants	6.73	6.78	V	3.53	3.61	J	3.2	3.17

REFERENCES

[1] N. Jakovljević, D. Mišković, M. Janev,D. Pekar, “A Decoder for Large Vocabulary Speech Recognition”, in *Proc. Int. Conf. on Systems, Signals and Image Processing IWSSIP,* pp. 1-4, Sarajevo, 2011.

[2] S. Ostrogonac, B. Popović, M. Sečujski, R. Mak, D. Pekar, “Language Model Reduction for Practical Implementation in LVCSR Systems”,in *Proc. Int. Scientific-Professional Symposium INFOTEH*, pp. 391-394, Jahorina, 2013.

[3] R. K. Moore, “A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners”, in *Proc. European Conf. on Speech Communication and Technology EUROSPEECH,* pp. 2582-2584, Geneva, 2003.

[4] N. Đurić, M. Sečujski, V. Delić, “Srpska SpeechDat(E) govorna baza snimljena preko fiksne telefonske mreže”,in *Proc. Conf. on Electronics, Telecommunications, Computers, Automation and Nuclear Engineering ETRAN*, pp. 333-336, Teslić, 2002.

[5] L. F. Lamel, J.-L. Gauvain, M. Eskénazi, “BREF, a Large Vocabulary Spoken Corpus for French”, in *Proc.European Conf. on Speech Communication and Technology EUROSPEECH*, pp. 505-508, Genoa, 1991.

[6] K. Itou, K. Takeda, T. Takezawa, T. Matsuoka, K. Shikano, T. Kobayashi, S. Itahashi, M. Yamamoto, “Design and Development of a Japanese Speech Corpus for Large Vocabulary Speech Recognition Assessment”, in*Proc. First Int. Workshop on East-Asian Language Resources and Evaluation,* pp. 98-103, Tsukuba,1998.

[7] S. Zablotskiy, A. Shvets, M. Sidorov, E. Semenkin, W. Minker, “Speech and Language Resources for LVCSR of Russian”, in *Proc.Int. Conf. on Language Resources and EvaluationLREC*, pp. 3374-3377, Istanbul, 2012.

[8] M. Sečujski, R. Obradović, D. Pekar, Lj. Jovanov, V. Delić,“AlfaNum System for Speech Synthesis in Serbian Language”, in*Lecture notes in computer science*, Vol. 2448, pp. 237-244, 2002.

[9] S. Ostrogonac, M. Bojanić, N. Vujnović-Sedlar, S. Suzić, “Detektor Štamparskih i Pravopisnih Grešaka za Srpski Jezik“ – Technical solution,http://www.ftn.uns.ac.rs/n1908637228/detektor-stamparskih-i-pravopisnih-gresaka-za-srpski-jezik--anspellchecker-, 2012.

[10] M. Sečujski, V. Delić, “A Software Tool for Semi-Automatic Part-of-Speech Tagging and Sentence Accentuation in Serbian Language”, in *Proc. Fifth Slovenian and First Int. Language Technologies Conf. IS-LTC*, pp. 226-229, Ljubljana, 2006.

[11] D. Pekar, R. Obradović, “C++ Library for Signal Processing – *slib*”, in *Proc. Telecommunications Forum TELFOR*, pp. 7.7:1-4, Belgrade, 2001.

[12] C. Gollan, H. Ney,”Towards Automatic Learning in LVCSR: Rapid Development of a Persian Transcription System”,in *Proc. Int. Conf. on Spoken Language Processing INTERSPEECH*, pp. 1441-1444,Brisbane,2008.