Published in:
Cybernetics and Information Technologies, vol.3, No.2, 2003, pp.101-108.

BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition over Telephone Channels

Atanas Ouzounov
Institute of Information Technologies, Sofia 1113
E-mail: atanas@iinf.bas.bg

Abstract: *In the paper is described the BG-SRDat (BulGarian language Speaker Recognition DATa) - a corpus in Bulgarian language collected over noisy analog telephone channels and intended for speaker recognition. The BG-SRDat comprises two different speech data, called Speech Data 1 (SD1) and Speech Data 2 (SD2), respectively. The SD1 is a reading text from a newspaper and its average length is about 40 seconds. The SD2 is a short phrase with length of about 2 seconds. The SD1 and the SD2 are uttered in various sessions by different number of speakers (male) – 26 and 13, respectively. To achieve more realistic real-world conditions the speech data is collected by different types of telephone calls (internal-routing, local and long-distance) and various acoustical environments (noisy offices, halls and streets). The main purpose of the BG-SRDat is to provide data for evaluation of various speaker recognition techniques with noisy telephone speech in Bulgarian language.*

Keywords: *Speech corpora, Speech databases, Speaker recognition.*

1. Introduction

It is well known fact that the standard speaker recognition databases (or corpora) can help researchers to compare the effectiveness of different approaches on common data and to allow them to select the more promising one.

In the last decade, a lot of speaker recognition databases for different languages are developed where the speech is recorded over telephone channels [1, 3, 5]. Based on the telecommunication standards in the countries where these corpora are designed, the calls are made mainly over the public switched telephone network (PSTN). By now this network has almost completely been made digital, except for the final connection to the subscriber. In this paper, we will call these databases standard databases.

It is known that now the large part of the telephone network in Bulgaria belongs to the old analog type. In this telephone network, we can observe a substantial amount of different type of noise (harmonic, impulsive, white) with variable level even in very short time period (typical noise is the impulsive noise due to the apparatus commutations). In addition, there are crosstalk, echo and time variability in the frequency response. These distortions make the speaker recognition more difficult task than this one with speech data passed via the digital telephone lines. When we develop and test speaker recognition algorithms with the standard databases we have to bear in mind that the obtained recognition performance always will be better than this one obtained with speech data collected over the analog telephone lines. That means if we develop an application, which will work with speech transmitted over the analog telephone channels, then it is reasonable to use in our research and this type of speech data.

We can consider the corpus with speech passed via the analog telephone lines as an addition to the standard databases. This additional corpus can help the researchers to evaluate the recognition techniques in more complicated channel environment.

2. Available speaker recognition corpora

At the present time the most popular publicly available speaker recognition corpora are: SIVA (Italian) [1, 3], PolyVar (Swiss-French) [1, 3, 5], POLYCOST (English and 13 other European languages – without Bulgarian)[1, 3, 5], KING-92 (English) [1], Switchboard I-II including NIST evaluation subset (English)[1, 6], OGI (English)[4] and AHUMADA (Spanish) [7]. We present only the corpora obtained by recording of the actual telephone calls. We do not consider these ones obtained by telephone channel simulation or by playing recorded speech into different type of handsets, etc.

In all mentioned above corpora the experiments are done under the next conditions:

-telephone type – various types of phones are used;

-telephone channels - PSTN or ISDN;

-acoustical environment at the speaker – mostly home or office.

At least a hundred speakers are included in the each of the mentioned above corpora (except for KING-92). All corpora included simultaneously male and female speakers spread approximately to the same proportion in each corpus (again except for KING-92). The type of speech material is highly varied – from fixed and prompted digit strings and read sentences to spontaneous speech. It is interesting to point out the POLYCOST corpus, which includes speech not only in English but and in speakers' native languages covering 13 European countries. More detailed description of the different corpora can be found in [1, 3, 7].

3. Description of the BG-SRDat

The BG-SRDat (BulGarian language Speaker Recognition DATa) is a speech corpus in Bulgarian language with small number of speakers recorded over noisy analog telephone channels. It can be used mainly for the fixed-text speaker verification and identification and in some cases for text-independent speaker identification. This corpus' release consists of speech files only and no transcription is included.

In our description we will focus on five factors: 1) type of speech (fixed-phrase, reading text, etc.); 2) number of speakers; 3) number and time separation of sessions; 4) channel, microphone and recording environment type; 5) files description.

3.1. Type of speech

The BG-SRDat comprises two different speech materials, called Speech Data 1 (SD1) and Speech Data 2 (SD2), respectively.

It is important to note that the SD1 and the SD2 are selected in such a way that their phonetic content complicates the speaker recognition process. To be more specific, we selected phonetically rich sentences in which the consonants dominate over the vowels.

The SD1 is a reading text from a Bulgarian newspaper (94 words in 4 sentences) and its average length is about 40 seconds. The English translation of the SD1 is:

> *"In his information Mr Gounev presents also his position about the work of the prosecutor's office after November 1st, 1990. He mentions that during that period prosecutor's office faced with rejection of all services to follow its instructions and recommendations. This put it in a position to appeal in front of persons and agencies for observance of the Law and to be criticized for not taking measures for preservation of the legality. Mr Gounev expresses his hope that the new legislation will make many discussions unnecessary and will assign the prosecutor's office the place, which it must occupy in a constitutional state."*

In the text below, we tried to show with symbols from the International Phonetic Alphabet (IPA) [12] the pronunciation of the Bulgarian text in the SD1.

"[V] [infor`matsiata] [si] [gospo`din] [`Gunev] [preds`tavia] [i] [sta`noviʃteto] [si] [za] [deino`sta] [na] [prokura`turata] [sled] [`pʌrvi] [no`emvri] [hi`liada `devetstotin i devetde`seta] [go`dina]. [Toi] [otbe`liazva], [tʃe] [prez] [`tozi] [peri`od] [prokura`turata] [se] [e] [`sblʌskvala] [s] [`otkaz] [na] [`vsitʃki] [rav`niʃta] [da] [`bʌdat] [`sledvani] [uka`zaniata] [i] [prepo`rʌkite] [i]. [`Tova] [ia] [e] [izp`ravialo] [v] [polo`ʒenieto] [da] [ape`lira] [pred] [li`tsa] [i] [`organi] [za] [`spazvane] [na] [za`kona] [i] [da] [`bʌde] [kriti`kuvana], [tʃe] [ne] [`vzema] [`merki] [za] [o`pazvane] [na] [zakonos`ta]. [Gospo`din] [`Gunev] [izra`ziava] [na`deʒda], [tʃe] [`novoto] [zakono`datelstvo] [ʃte] [nap`ravi] [`mnogo] [`sporove] [iz`liʃni] [i] [ʃte] [otre`di] [na] [prokura`turata] [`miastoto], [ko`eto] [tia] [`triabva] [da] [za`ema] [v] [`pravovata] [dʌr`ʒava]."

That way we represented roughly the authentic sound of the Bulgarian phonemes. To show the exact English pronunciation of the Bulgarian phonemes included in this text is beyond the scope of the author and did not include in the paper.

The SD2 is a short phrase in Bulgarian language and its average length is about 2 seconds. The phrase is (with Latin letters): *"Zdravei Manolov. Kak se chuvstvash dnes?"*. Its English meaning is "*Hello Manolov! How are you?*". The pronunciation (roughly) is – *"[zdra`vei:] [ma`nolov]! [kak] [se] [`tʃuvstvaʃ] [dnes]?"*.

3.2. Number of speakers

The corpus contains speech data from 32 male speakers in the age between 26 and 45 years. The text material in SD1 is read out by 26 speakers while the phrase in SD2 is uttered by 13 speakers. Seven of speakers have records of both speech data.

3.3. Number and time separation of sessions

3.3.1. SD1

There are at least 2 sessions per speaker in SD1. Only one record of the reading text belongs to each session. The period between sessions is about 3 months.

3.3.2. SD2

At least 10 sessions per speaker are included in the SD2 and usually a pair of utterances belongs to each session. In every session, the data are collected on the same day from different telephone calls made from the same calling number (i.e. from the same telephone apparatus). The telephone calls in different sessions are from different telephone numbers located in different districts in the city of Sofia. The period between sessions is about a week.

3.4. Channel, microphone and recording environment type

Only the first session in the SD1 is recorded in a quiet office with a high quality microphone. All the rest sessions in the SD1 and the SD2 are recorded over the telephone channels.

3.4.1. Handset type

The speakers used mainly two type of telephone handset – with carbon and with electret microphone.

3.4.2. Acoustical environment

In attempts to include in our corpus various real-world acoustical environments, we recommended the speakers to make the calls from many different places, e.g., quite and noisy

offices, halls and pay phones in the streets. It is worth noting that more then a half of telephone calls are made from pay phones placed in the noisy streets.

3.4.3. Type of telephone call

The telephone calls are internal-routing, local and long-distance ones. There are some exceptions, namely few calls are made from cellular phone and few utterances are recorded directly from the microphone in the telephone handset (without channel transmission).

3.4.4. Recording conditions

The telephone calls are recorded on a tape deck. The first session in the SD1 is recorded in a quiet office with a high quality microphone UHER M534.

Later the audio records are digitized at 8 kHz on 16 bits (mono, PCM), after low-pass filtering at 4 kHz. Each utterance is saved in a separate file in WAV format. No additional processing of the digitized speech is applied.

3.5. Files description

The file name for each record is formed in the following way: the file name is *'s_xxyyzz.wav'* , where:

- *s* - the speech data type SD1 or SD2, *'1'* and *'2'*, respectively;
- *xx* - the identification number of the speaker (ID), starts from *'01'*;
- *yy* - the number of the session, starts from *'01';*
- *zz* - the number of the utterance in the particular session, starts from *'01'.*

For example, the file name *'2_110501.wav'* means that this is a record of speech data from type SD2, for speaker with ID=11, the fifth session and the first utterance in this session. The file structure on the disk is *'speaker#/speech_data#/session#/utterance_name.wav'* or for the mentioned above file name the path is *'11/SD2/05/2_110501.wav'*. For more clearness, we placed the files in different directories, but since the file names are unique, it is possible to put them in one common directory.

We did not divide our speech data into two groups – for enrolment and for recognition – as it is done in YOHO database [2]. The reason is that the utterances among the sessions are with different channel and noise characteristics. We leave the researcher to decide which sessions will be used for enrolment and which ones for recognition purposes. In order to help the search among speech data each utterance is described with a few parameters in a so-called Utterances Description Table (UDT). Each speaker has own UDT (for all his utterances). If a particular speaker has utterances in the SD1 and the SD2 then for him two UDTs will be designed. The following parameters are included in the UDT:

-speaker ID;
-used speech data set – SD1 or SD2;
-type of the headset microphone;
-position of the speaker;
-telephone calls type;
-noise presence marks.

Based on simple auditory tests the noise presence marks are used to note the presence (just yes or no) of few basic forms of noise in speech data. These few forms of noise are divided quite conditionally into two groups. The first one embraces noises mainly due to the equipment operation: impulsive noise, background noise (e.g., white, harmonic) and crosstalk. The environmental noises generated at the speaker position (e.g., background conversations, music and traffic noise) belong to the second group. It is obvious that the "noise presence mark" is a subjective parameter and it is included in the UDT mainly for illustrative purpose. The UDT for speaker with ID=25 and SD1 is shown in the Table 1. The UDT's parameters are noted as following:

-hifi mic – dynamic hi-fi microphone (it is used for direct recording on the deck without telephone call);
-carbon – telephone handset's carbon microphone;
-local – telephone call via a local exchange;
-office/street – the place from where the speaker makes the call;
-BN – background noise;
-PN – impulsive noise;
-CT – crosstalk;
-BC – background conversation;
-M – music;
-TN – traffic noise;
-*/- – noise presence mark (yes/no);

Table 1. Utterances Description Table (UDT) for the SD1 and speaker ID=25

Speech Data 1 Speaker ID = 25											
№	***File name***	***Session No./ Utterance No.***	***Handset micro-phone***	***Speaker position***	***Phone call***	***Noise presence marks***					
						Equipment operation noise			***Environmental noise at speaker position***		
						BN	***PN***	***CT***	***BC***	***M***	***TN***
1	1_250101.wav	1/1	hifi mic	office	no call	-	-	-	-	-	-
2	1_250201.wav	2/1	carbon	office	local	*	*	-	-	*	-
3	1_250301.wav	3/1	carbon	street	local	*	*	*	-	*	*
4	1_250401.wav	4/1	carbon	office	local	*	-	-	-	-	-
5	1_250501.wav	5/1	carbon	office	local	*	-	-	-	-	-

In Figs. 1 and 2 are shown the waveforms and the narrowband spectrograms of two wav files (belonging to different speakers) from the SD2. The telephone noise in these files is typical of analog lines. It is a mixture of a background noise and some kind of impulsive noise (crackles, pops, etc.).

As can be seen in Fig.1 the dominated noise in the first file is an impulsive noise (crackles). The amplitudes of this noise often are higher than those ones of the speech signal - see waveform in Fig.1 (a). Furthermore, between time axis tick 4.7 s and the end of the file - see Fig.1 (b), can be detected segments with crosstalk (low-level harmonic signal).

Along the length of the second wav file are clearly noticed pops and crosstalk (induced busy line signal) – see spectrogram in Fig.2 (b).

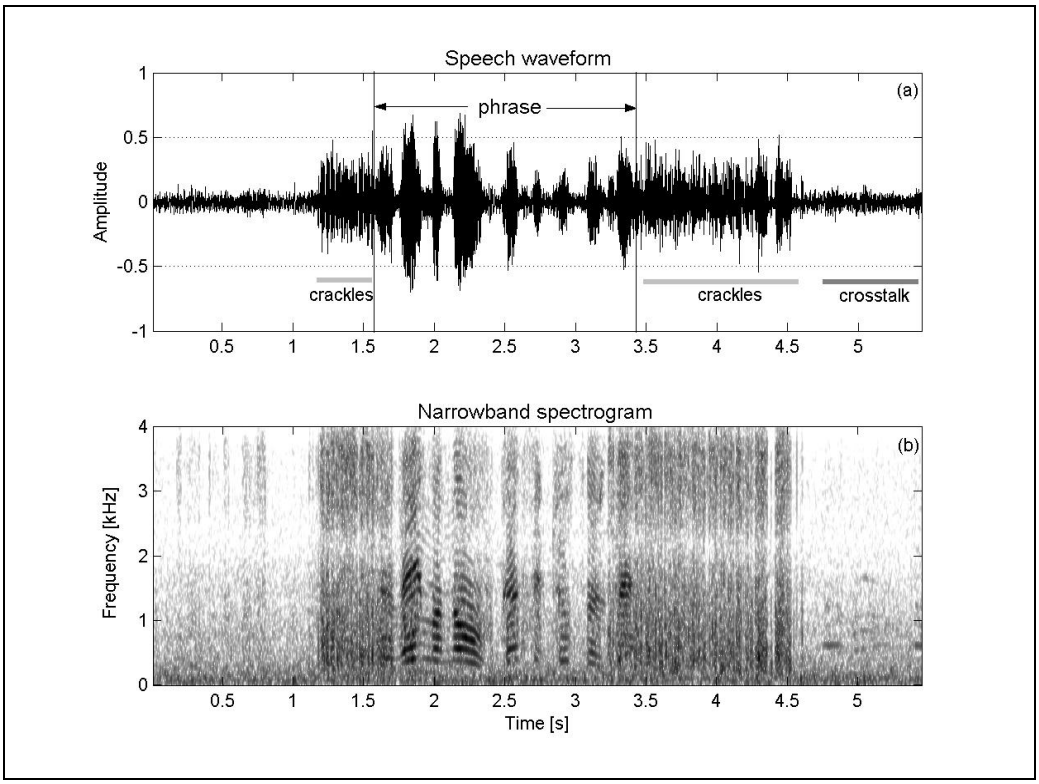

Fig.1. File from SD2 with crackles.

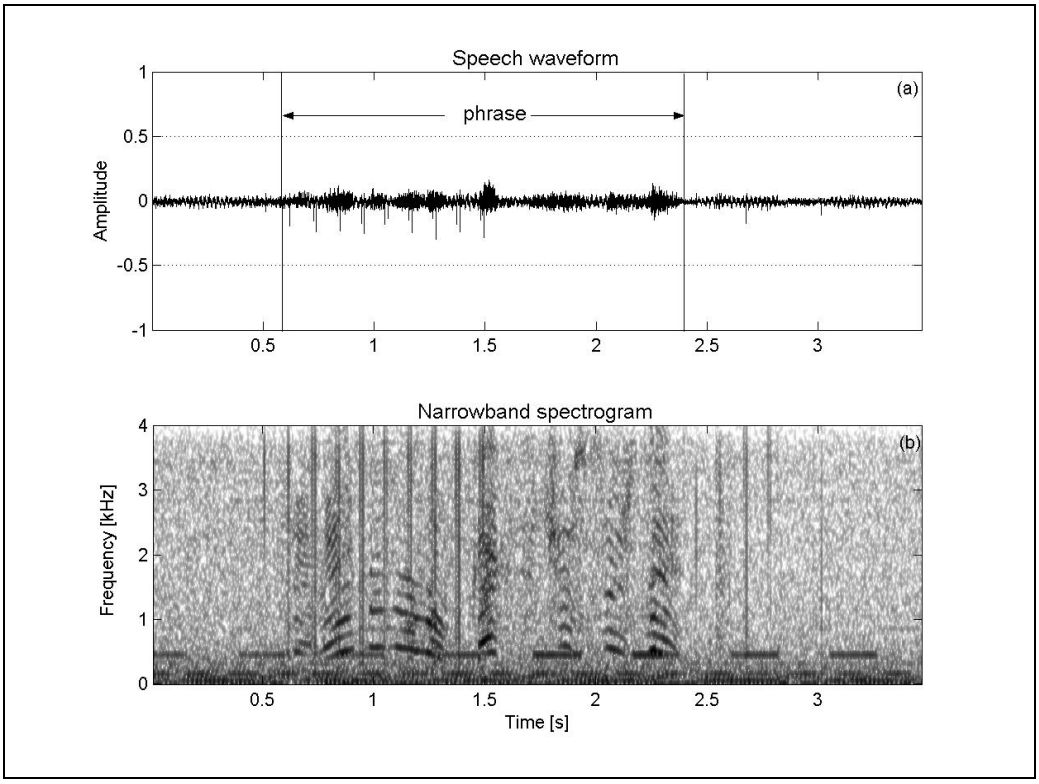

Fig.2. File from SD2 with pops and crosstalk.

4. Using of BG-SRDat

The BG-SRDat is utilized in various experimental setups for fixed-text speaker verification and identification, text-independent speaker identification and speech segmentation tasks [8, 9, 10, 11].

In the fixed-text speaker identification tasks, generally we used the first four or five sessions of SD2 as enrolment data and the rest sessions – as identification data. During the training procedure we did not make any selection among the utterances intended for reference creation, i.e., the multi-style training is applied. That means the reference can be created from a clean utterance together with a very noisy one. Because there are essential difference among sessions in channel noise and acoustical environment then the choice of the enrolment and recognition data strongly affects on the final recognition performance [11].

In the text-independent speaker identification tasks the parts of utterances in the sessions different from the first one in the SD1 (the first session is not telephone speech) are used as training data and the rest data (with different linguistic content compared to the training one) from the rest sessions - as identification data [9, 10].

An additional clean speech session in the SD1 and the longer duration of the reading text allowed us to develop and evaluate some approaches for speech segmentation (e.g., voiced-unvoiced-silence segmentation) [8].

5. Conclusions

The BG-SRDat integrates two separate sub-corpora (SD1 and SD2) collected over telephone lines with different purposes and in a different time. The first sub-corpus (SD1) is utilized for text-independent speaker identification as well as for a voiced-unvoiced-silence speech segmentation. The second sub-corpus (SD2) includes one short phrase and it is intended only for fixed-text speaker recognition. Since the data in these corpora are collected independently due to coincidence, only part of speakers has records in both of them.

Now we consider that the BG-SRDat is an incomplete corpus. To improve it, the number of speakers needs to be increased and more varied speech material should be added. Moreover, the speech files transcription should be included in the next corpus release.

REFERENCES

1. Campbell, J., D. Reynolds, Corpora for the evaluation of the speaker recognition systems,- ICASSP’99, pp.829-832.
2. Campbell, J., Testing with the YOHO CD-ROM voice verification corpus, - ICASSP’95, pp.341-344.
3. Melin, H., Databases for Speaker Recognition: Activities in COST250 Working Group 2, *COST 250 - Speaker Recognition in Telephony, Final Report 1999*, European Commission DG-XIII, Brussels, August 2000.
4. Cole, R., M. Noel, V. Noel, The CSLU Speaker Recognition Corpus, - ICSLP’98, vol. 7, pp.3167-3170.
5. Evaluations and Language resources Distribution Agency (ELDA), Telephone Language Resources, http://www.elda.org/article17.html.
6. National Institute of Standards and Technology (NIST), Speech Group, Speaker Recognition Evaluation Data, http://www.nist.gov/speech/tests/index.htm.

7. Ortega-Garsia, J., J. Gonzalez-Rodriguez, V. Marrero-Aguiar, AHUMADA: A large speech corpus in Spanish for speaker characterization and identification, -*Speech Communication*, vol. 31, 2000, pp. 255-264.
8. Ouzounov, A., E. Peev, Speech Segmentation using Neural Networks, - *Problems of Engineering Cybernetics & Robotics,* No.45, 1996, pp.3-12.
9. Ouzounov, A., Text-Independent Speaker Identification using a Hybrid Neural Network and Conformity Approach, - In: Proceeding of the IEEE International Conference on Neural Networks, Houston, USA, 1997, vol. IV, pp.2098-2102.
10. Ouzounov, A., Speaker Identification using a Hybrid Neural Network and Conformity Approach, - In: Proceedings of The First European Conference on Signal Analysis and Prediction, Prague, 1997, pp.455-459.
11. Ouzounov, A., An Evaluation of DTW, AA and ARVM for Fixed-Text Speaker Identification, - *Cybernetics and Information Technologies*, vol. 3, No.1, 2003, pp.3-10.
12. The Pocked Oxford Dictionary, Clarendon Press, Oxford, 1984.