БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

КИБЕРНЕТИКА И ИНФОРМАЦИОННИ ТЕХНОЛОГИИ · Том **3**, № 2
CYBERNETICS AND INFORMATION TECHNOLOGIES · Volume **3**, No 2

София . 2003 . Sofia

Applications

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

Atanas Ouzounov

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

Abstract: *The paper describes the BG-SRDat (BulGarian language Speaker Recognition DATabase) – a corpus in Bulgarian language, recorded over noisy analog telephone channels and intended for speaker recognition. The BG-SRDat comprises two separated speech corpora, 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 acoustical environments (noisy offices, halls and streets). The BG-SRDat purpose is to help the researchers to evaluate various speaker recognition techniques for noisy telephone speech in Bulgarian language and to select the more promising one.*

Keywords: *Speech corpora, speech databases, speaker recognition.*

1. Introduction

It is known that the standard speaker recognition databases (corpora) permit the researchers to compare the effectiveness of different approaches on common data and to help 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 our paper we will call these databases standard databases.

It is known that at the present time the greatest 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 (harmonic, pulse, white) of noise with variable level even in a very short time period (typical noise is the pulse noise due to the apparatus commutations). There are also crosstalks, echo and time variability in the frequency response. These disturbances make the speaker recognition with speech data collected over them a more difficult task than this one with speech data passing through 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 will always be better then this one obtained with speech data collected over the analog telephone channels. That means that if we wish to develop a real-world application that will work with speech passed through the analog telephone channels and if we wish to have a realistic evaluation of the recognition performance then we have to use in our research work namely this type of speech data.

We can consider the speaker recognition corpus with speech passed through the analog telephone lines as an addition to the standard databases. This additional database 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 don't 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 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 DATabase) is a speech database 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.

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 speech materials (i.e. two separated speech corpora), 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 contents makes difficult the speaker recognition process. To be more specific, we selected texts in which the consonants are more then 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 mister 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. Mister 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 bellow we try 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] [`p∧rvi] [no`emvri] [hi`liada `devetstotin i devetde`seta] [go`dina]. [Toi] [otbe`liazva], [tʃe] [prez] [`tozi] [peri`od] [prokura`turata] [se] [e] [`sbl∧skvala] [s] [`otkaz] [na] [`vsitʃki] [rav`niʃta] [da] [`b∧dat] [`sledvani] [uka`zaniata] [i] [prepo`r∧kite] [i]. [`Tova] [ia] [e] [izp`ravialo] [v] [polo`jenieto] [da] [ape`lira] [pred] [li`tsa] [i] [`organi] [za] [`spazvane] [na] [za`kona] [i] [da] [`b∧de] [kriti`kuvana], [tʃe] [ne] [`vzema] [`merki] [za] [o`pazvane] [na] [zakonos`ta]. [Gospo`din] [`Gunev] [izra`ziava] [na`dejda], [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] [d∧r`java]."

That way we represent 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 didn't 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 SD1 and the SD2 are pronounced by different number of speakers, but all recorded speakers are men in the age between 26 and 45 years. The SD1 is uttered by 26 speakers (named Set 1 – S1) and the SD2 by 13 speakers (named Set 2 – S2). Only seven speakers have utterances in both speech data.

3.3. Number and time separation of sessions

3.3.1. SD1

For the SD1 there are at least 2 sessions per speaker (14 speakers from S1 have at least 3 sessions). Only one record of the reading text belongs to each session. The time period between sessions is about 3 months.

3.3.2. SD2

For the SD2 there are at least 10 sessions per speaker (maximal number of session is 15) and only a pair of utterances belongs to each session. In every session the records are done by different telephone calls but from the same calling number (i.e. from the same telephone apparatus) and the time period between records is a day. The telephone calls for different sessions are from different telephone numbers located in different districts in the city of Sofia. The time period between sessions is a week.

3.4. Channel, microphone and recording environment type

Only the first session for the SD1 is recorded in a quiet office with a high quality microphone. All the rest sessions for 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 electret microphone.

3.4.2. Acoustical environment

In attempts to include various real-world acoustical environments in our corpus 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 done from pay-phones (without booth) placed in the noisy streets.

3.4.3. Type of telephone channel

The speakers are made mainly the internal-routing, local and long-distance calls. 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. Only the first session of the SD1 is recorded in a quiet office with a high quality microphone UHER M534.

Later on, these audio records are digitized at 8 kHz on 16 bits (mono, PCM), after low-pass filtering at 4 kHz. For that purpose we utilized the audio board ProAudio Spectrum 16 by Media Vision Inc. 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 of the disk is *"speech_data#/speaker#/ session#/utterance#.wav"* or for the mentioned above file name the path is *"SD2/ 11/05/2_110501.wav"*. We placed the files in different directories for more clearness, but since the file names are unique it is possible to put them in one common directory (or in separated speaker's directories), if it is necessary.

We did not divide our speech data in two groups – for enrollment and for recognition – as it is done, e.g., 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 enrollment and which ones for recognition purposes. Each utterance in the database 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 parameters in the UDT are:

– speaker ID;

– speech data set – SD1 or SD2;

– type of the microphone;

– the position of the speaker;

– type of the channel;

– channel properties:

  – background noise (white, harmonic);

  – pulse noise;

– crosstalk;

– environmental noise generated at the speaker position:

– background conversations;

– music;

– traffic noise;

In the UDT we marked only the presence of the mentioned above disturbances (as a result of the auditory tests). It is clear that the parameters included in the UDT are very simple to describe the utterance characteristics. They can be utilized only for preliminary utterance selection.

In Table 1 the UDT for speaker with ID=22 and SD1 is shown.

Table 1. Utterances Description Table (UDT) for the speaker with ID=22 and SD1

Speaker Data 1 Speaker ID=22											
№	Session	Utter-ance	Micro-phone	Speaker position	Channel type	Channel properties			Noise at speaker position		
						BN	PN	CT	BC	M	TN
1	1	1	dynamic	office	directly	–	–	–	–	–	–
2	2	1	carbon	hall	local	y	–	–	–	–	–
3	3	1	carbon	street	local	y	–	y	–	–	–
4	4	1	carbon	street	local	y	y	–	–	–	y

In Table 1 we used the following abbreviations:

– dynamic – dynamic hi-fi microphone;

– carbon – telephone handset's carbon microphone;

– local – telephone call via a local exchange;

– directly – no telephone call, the utterance is recorded directly from the microphone on the deck;

– BN – background noise;

– PN – pulse noise;

– CT – crosstalk;

– BC – background conversation;

– M – music;

– TN – traffic noise;

In Fig. 1 three typical wav files from the SD2 are shown. They belong to the three different speakers. As can be seen in the Figure 1 the signal-to-noise ratio in the records is low and the pulse noise amplitude often is higher than the amplitude of the speech signal.

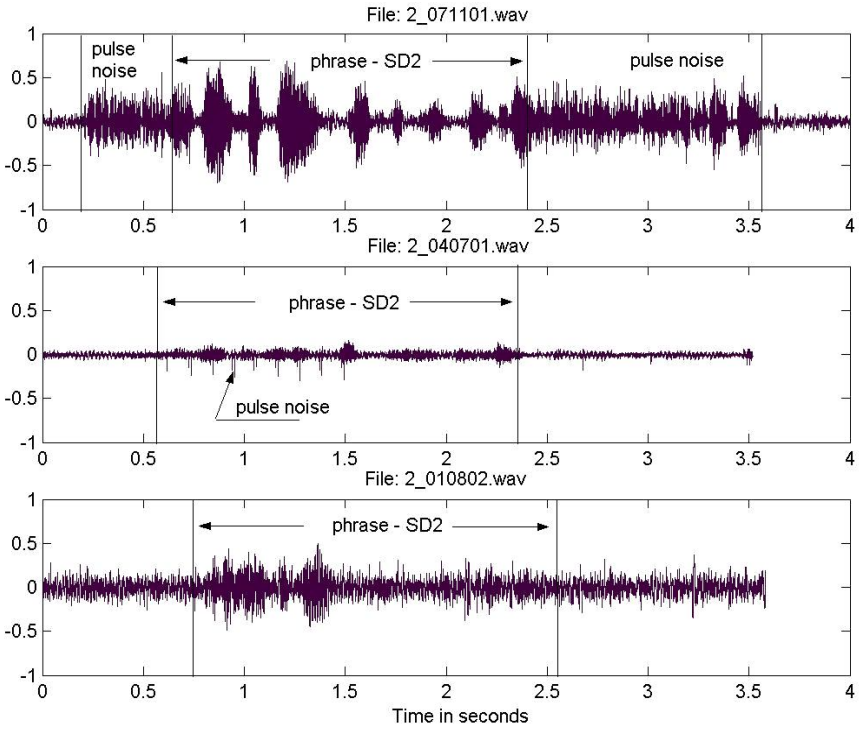

Fig.1. Three typical wav files from SD2

4. Using of BG-SRDat

The BG-SRDat is utilized by the author 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 4 or 5 sessions of SD2 as enrollment 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 enrollment and recognition data strongly affects on the final recognition performance [11].

In the text-independent identification tasks the first 10 or 20 s from utterance in the second session of 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

In fact the BG-SRDat integrates two separate corpora (SD1 and SD2) collected over telephone lines with different purposes and in a different time.

The first speech corpus (SD1) is utilized for text-independent speaker identification as well as for a voiced-unvoiced-silence speech segmentation. The second corpus (SD2) includes only one short phrase and it is intended only for fixed-text speaker recognition.

Since the speech data in these corpora are collected independently, by reason of coincidence only part of speakers has records in both of them.

Currently we consider that the BG-SRDat is an incomplete corpus. It is necessary to increase the number of speakers (have to be at least 100), to include female speakers (typically a half of speakers in database) and to include data passed via other communication channels, e.g. cellular phones, VoIP, etc.

References

1. Campbell, J., D. Reynolds. Corpora for the evaluation of the speaker recognition systems. – ICASSP'99, 829-832.
2. Campbell, J. Testing with the YOHO CD-ROM voice verification corpus. – ICASSP'95, 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, **7**, 3167-3170.
5. http://www.elda.fr/catalogue/tabsp1.html
6. http://www.nist.gov/speech/tests/spk/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, **31**, 2000, 255-264.
8. Ouzounov, A., E. Peev. Speech segmentation using neural networks. – Problems of Engineering Cybernetics and Robotics, **45**, 1996, 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, 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, 455-459.
11. Ouzounov, A. An evaluation of DTW, AA and ARVM for fixed-text speaker identification. – Cybernetics and Information Technologies, **3**, 2003, No 1, 3-10.
12. The Pocked Oxford Dictionary. Oxford, Clarendon Press, 1984.

BG-SRDat – база говорен материал на български език, записан по телефонен канал и предназначен за разпознаване на диктори

Атанас Узунов

Институт по информационни технологии, 1113 София
E-mail: atanas@iinf.bas.bg

(Р е з ю м е)

В работата е описана BG-SRDat (BulGarian language Speaker Recognition DATabase) – база от говорен материал, записан по зашумени аналогови телефонни линии и предназначен за разпознаване на диктори. BG-SRDat включва два отделни говорни материала – SD1 и SD2. Първият материал е текст, публикуван във вестник, а вторият е фраза от непринуден разговор. Средната продължителност на произнасянията за първия материал е около 40 s, а за втория – около 2 s. Произнасянията на двата говорни материали са реализирани в няколко последователни сесии и с различен брой диктори-мъже, съответно 26 за SD1 и 13 за SD2. Записани са различни по вид телефонни разговори – вътрешни (чрез учрежденска телефонна централа), градски и междуградски. Дикторите са използвали телефонни апарати, разположени на различни места – офиси, зали и улици. Основната цел при създаването на BG-SRDat е тя да съдържа данни на български език, получени при реални условия и по този начин да помогне на изследователите при разработка на алгоритми за разпознаване на диктори.