Design and implementation of a Russian telephone speech database[1]

Kouznetsov Vladimir

Moscow State Linguistic University, Ostozhenka 38, Moscow, Russia, 119837,

Chuchupal Vladimir[2], Makovkin Konstantin, Chichagov Alexander

Computer Center, Russian Academy of Science, Vavilova 40, Moscow 117967, RUSSIA

Abstract: This paper presents a Russian spoken language database. The speech material design comprises two types of speech data and data collection protocols. The speech signal was captured by microphone and telephone card simultaneously. To record more than one hundred of speakers over public telephone network, an original T-1 based telephone data collection system have been developed and implemented.

1**. INTRODUCTION**

The objective of creating a Russian spoken language database was two-fold. On the one hand, the data base was meant to provide the researcher with a basic amount of speech material for general speech research in the field of telecommunication. On the other hand, the speech corpus was designed to obtain a controlled amount of speech for training conventional context-dependent phone models, that may be used for the development of commercial speech recognition engines. It is also planned to use the database in the development and testing of the speech processing algorithms for noise cancellation, channel equalisation and in assessment of system performance. For research purposes telephone signals were simultaneously recorded with 'clean' speech pressure signal.

2. SPEECH CORPRA DESIGN

The speech material design comprises two types of data collection protocols.

The first one is aimed at creating a representative corpus of spoken Russian - around 500 sentences are read by 10 speakers (5 men and 5 women). The speech material is manually segmented and phonetically labelled.

In this set of the sentences each phoneme occurs at least ten times. In the Sound Frequency Table the frequency of occurrence of phonemes in the sentences is presented. Transcription of the sentences was carried out using a modified version of the program developed in MGU under supervision of Dr. O. Krivnova. Using Latin alphabet and ASCII symbols for transcription of Russian speech we aimed at a greater portability of the segmentation data. In most cases there is a direct correspondence between Cyrillic and Latin transcriptions. The symbol '^' is used to mark stressed vowels; combination of a vowel sign with figures '0' and '1' means that the vowel is strongly reduced and is found in pre- or poststressed position, correspondingly. Phonemes /a, e, o, i/ when strongly reduced after palatalised consonants in pre- and poststressed position are designated as W0 and W1, correspondingly. Vowel clusters are used to describe cases when the vowels are as a rule non-segmentable. The apostrophe marks palatalised consonants. Semicolon denotes long phonemes. In the process of labelling speech segments the basic transcription symbols could be specified by the following modifiers: '~' –for nasalised sounds; 'DV' – for devoiced sounds; 'V' – for voiced ones; 'CR' marks a creaky sound.

Several transcription signs are designed to capture non-verbal sounds and pauses.

The objective of the second data collection protocol is to elicit task specific tokens of read and spontaneous speech from about 100 speakers. The protocol includes prompts, which elicit a total of 8 to 10 minutes of speech from each caller. The responses fall into four categories: (a) answers to requests for specific information; (b) reading of a set of phonetically rich sentences and a list of numbers of different length; (c) continuous speech on selected topics; (d) extemporaneous speech. The speech material is accompanied by non-time-aligned phonetic transcription.

3. DATA FORMAT.

Telephone speech signal (including cellular type) is recorded simultaneously with 'clean' speech pressure signal. The 'clean' data allows making various insightful comparisons at the stage of research and development of the noise enhancement algorithms.

For the 'clean' signal the data format is 22050 Hz., 16-bit mono, Microsoft WAV, for the telephone channel recordings - mu-law and 8-bit resolution format. The labelling and segmentation data are stored in the format that

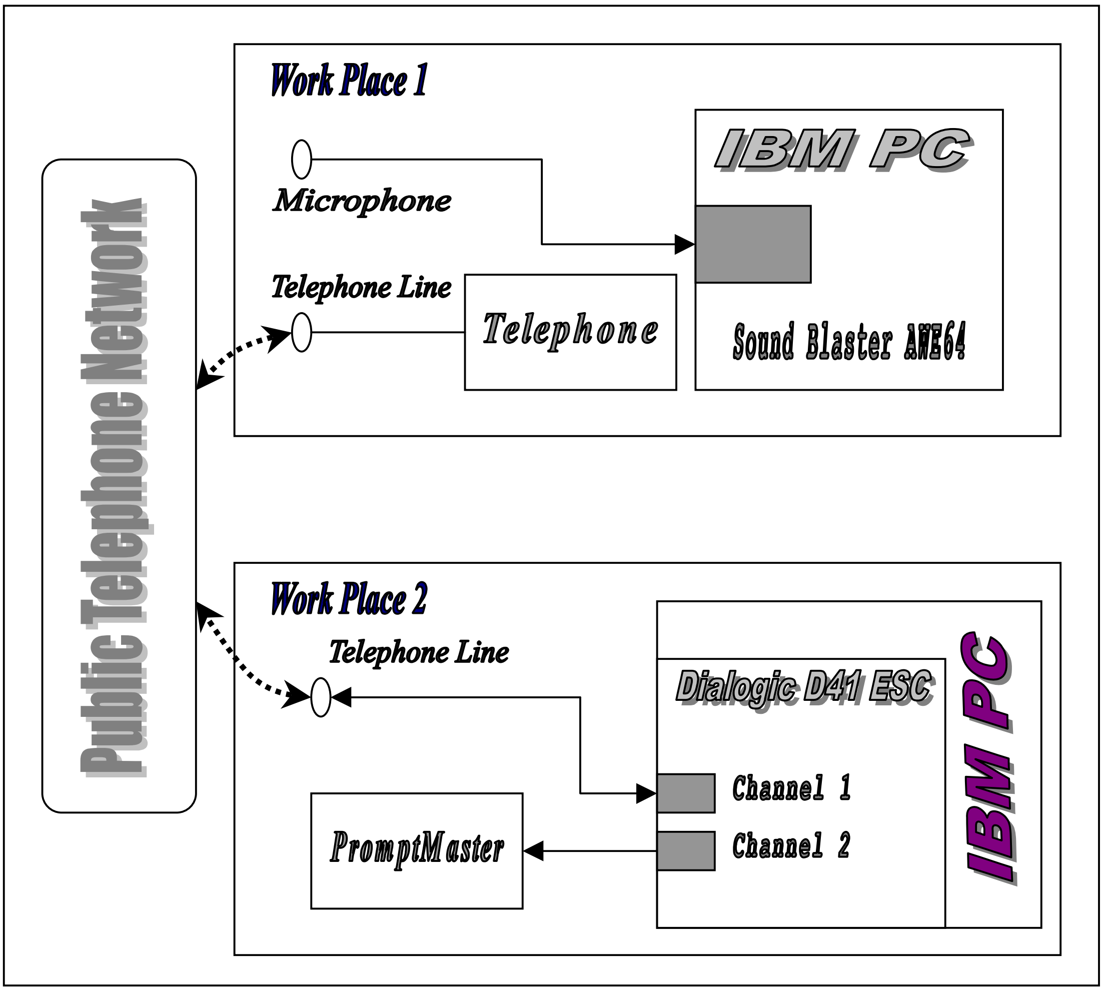
[1] The work was supported in part by Russian Fund of Fundamental Research, grant №98-01-00561

[2] Correspondence author chuchu@ccas.ru

is compatible with that of the TIMIT database, thus solving the problem of the database availability and dissemination.

4. DATA COLLECTION SYSTEM

To record more than one hundred of speakers, an original T-1 based telephone data collection system have been developed and implemented. The data collection hardware platform is an Intel Pentium PC with a four-channel Dialogic D41 telephone board. The figure below shows the data collection hardware and software configuration.



Data capture was performed at the two work places simultaneously.
To control recordering session the special application has been developed. The program runs under Windows NT version 4.0 and performs the following operations.
For the recording of the first part of data base (the phonetic balanced sentences) the semi-automatic protocol is implemented.

1. At the start of the session the operaton launchs the application from the workplace2 and set up the directories for the data files. The the program is waiting for the incoming call (if nessecary he program from workplace2 can initiate the session itself);
2. After the connection with the workplace2 is established the D41 card channels are in the duplex mode so that the operator at workplace2 can use the PromptMaster telephone line the usial phone to negotiate with the calling party (at workplace2). It was useful since the first tests were rather long (40-50 minutes).
3. When preliminary negotiations with the other party is finished the record mode is set. The D41 start record the incoming signal into files. The D41 board channels switches so that the PromptMaster was in the half-duplex mode. The operator still could listen the record options and quality of signal, however no any extra sound from the receiving party could distort the signal.
4. When test is finished or at some unpredictable curcumstances the operator switches the D41 into full duplex mode and then stop recording or negotiate with the calling party.

For recording the second part of the data base the other full automatic mode of data capture has been used.
The data collection application software allows to use the text scripts that describes the sequence of operations to establish connection and go through the test. The program is running as the script-driven dialog program.
For this case the following protocol is supported.

1. At the session start the operator at the workplace2 launchs the application and enters the name of the script.

   Then application is switched into waiting mode. If necessary the operation could enter the other party's telephone number and initiate the call from the workplace2.
2. When connection is established the script is launched automatically. The D41 channels are in the half-duplex mode, so if somebody (i.e. operator) if present he/she could listen what is going on and control recording options manually.
3. The data recording is run in automatic mode. The calling party at the first workplace is driven by the prompts and questions from the data collection software. When all tests are passed program disconnects automatically.

The recording script is actually the sequence of ASCII strings. Each string describes the separate operation that should be performed by the data collection software. For example, the following strings define the simple prompt-driven record operation.

PROMPT *filename_prompt.vox*
CAPTURE *filename_capture.vox* *SILENCE, DIGIT_5, OPERATOR*

The result of execution of above script is that the file *filename_prompt.vox* will be played into the telephone. Than program starts recordering of signal into the file *filename_capture.vox.* The record operation will be finished if dictor will keep silence (during 5 seconds or more), or the digit number 5 (dual tone multifrequency) will be pressed. Also the operator can stop the record procedure himself, by pressing the STOP push button .

SOUND FREQUENCY TABLE

SOUNDS	**FREQUENCY**	**SOUNDS**	**FREQUENCY**	**SOUNDS**	**FREQUENCY**
A	736	G	110	W0A	16
A^	444	G'	57	W1	452
A0	195	JH	10	J'	162
A0A	5	K	316	J':	1
A1	382	K'	80	B	133
E	38	K:	1	B'	66
E^	344	L	371	B:	1
E0	2	L'	319	CH’	135
E1	17	M	169	D	191
I	444	M'	83	D'	116
I^	282	M:	2	D':	1
O	9	N	383	D:	5
O^	381	N'	227	DZ	18
U	136	N':	2	F	100
U^	110	N:	14	F'	57
U0	16	P	270	T:	2
U1	49	P'	73	TS	147
Y	58	P':	1	TS:	3
Y^	104	R	466	V	232
Y0	10	R'	185	V'	114
Y1	107	R:	3	V:	3
JA	14	S	313	X	84
JA1	17	S'	168	X'	44
JE	14	S':	3	XV	17
JE1	61	SH	119	Z	177
JI	15	SH'	50	Z'	79
JU	18	SH:	1	Z:	3
JU1	27	T	453	ZH	85
W0	115	T'	155	ZH:	8