

O.F. Krivnova
THE SCOPE OF SPEECH CORPORA APPLICATION AND EXPERIENCE
OF THEIR DEVELOPMENT

Moscow State Lomonosov University, Philological Faculty
Russia, 119899 Moscow, Vorobyevy gory, I gum. korpus
Tel.: (495) 939-26-01
E-mail: okri@philol.msu.ru

Speech corpora, also called speech databases, is an important type of language recourses. A corpus often contains computer programs, which provide creation, organisation and management of language recourses proper, including phonetical ones. The interest in the creation of speech corpora was initiated to a large extent by works in the field of automatic speech recognition, where researchers have to deal with a great acoustic variability of sound units, which has very different sources, ranging systematic contextual variability, caused by coarticulation, a speaker's psychological state or technical parameters of a microphone used for the recording of speech material. Modern recognition systems are usually trained on very large amounts of sounding speech, recorded from many speakers (not less than 100). In the last decade there has been a noticeable switch from "manual" rules and algorithms to a corpus modelling in automatic speech synthesis. It is especially important for the modelling of prosodic characteristics of speech, its emotional content and expression and also for the imitation of individual peculiarities of a speaker's voice. Speech corpora by themselves are of scientific interest, and they are necessary for a lot of scientific tasks connected with analysis and description of speech in various languages. The report presents an analysis of main spheres of speech corpora application, there is also a brief summary of the experience in their development, including the one conducted on Russian material.

1. Speech corpus as a variety of language recourses. Speech corpora, also called speech databases, is an important type of language recourses. The latter term is usually used to signify any, often large, sets of linguistic data and descriptions, represented in electronic format and specially organised for the development, improvement and evaluation of systems and algorithms of processing speech and language material in technological applications.

1. **Speech corpus** is a structured aggregate of speech fragments, provided with programmed means of their access. **Speech fragment** as a basic corpus unit is a digital fragment of speech signal, accompanied with associated information of a certain type(s). Nowadays, the task of creating large, diverse and informationally "rich" (multi-level) speech corpora, together with a convenient and reliable set of tools for their development and usage, is becoming more and more urgent both for computer applications and fundamental phonetic researches. Modern systems of speech recognition with highest degree of reliability are mainly based on methods of statistic modelling of speech and language phenomena and require training on big amounts of annotated speech data, recorded from many speakers (not less then 100 people).

Modern approach to text-based speech synthesis founded on concatenation of acoustic fragments of various limentions, also presupposes the use of large speech corpora [1]. Experts [3] regard corpus-based approach as determining for the development of synthesis technologies, especially for modeling of prosodic speech characteristics and speaker's individual features. They also point out such advantages of this approach as formalisation of training procedures, the use of iterative training process with correction of new and controlled mistakes, the possibility of control and objective evaluation of various applied systems on standard basis (for the same speech corpora). Experience shows that, on the condition of availability of speech corpora and training technology, the creation of a prototypical version of automatic speech recogniser or synthesizer does not require such a long time. As it is said in specialized literature, it takes from two to six months [2]. That is quite important for commercially-oriented projects.

It would be a mistake to think that speech corpora are of interest only for the development of speech technologies. The use of representative speech corpora, annotated with special information, the level of the development of speech technologies and constantly increasing power of computers' power provide scientists with an earlier non-existent opportunity for conducting large-scale and statistically reliable phonetic researches on various speech material.

2. From the history of speech corpora. The first speech corpora appeared in mid-1980s in the USA, where their development was mostly sponsored by the Ministry of Defence. With its support, there were created: TI-DIGITS corpus (1984) to test systems of recognition of digits and digital sequences; Road Rally – to analyse and recognise key words (word spotting), and King Corpus for systems of speaker recognition. Within a state programme of the development of linguistic technologies, known as ARPA/DARPA (the Advanced Research Projects Agency), the same ministry financed the creation of a famous American corpus TIMIT (1980-1990), which served as a prototype for many other speech corpora. Thanks to the same financial support, there were developed specialised speech corpora – Recourse Management (RM) and Wall Street Journal (WSJ) for researches in the field of continuous speech recognition, as well as Air Travel Information Service (ATIS) for studying spontaneous speech and understanding of natural language in dialogue systems.

Experience has shown that the creation of a qualitative speech corpus is a rather complicated technological task, requiring considerable financial and personnel investments. Problematic points in this task are still financial maintenance, the necessity of team work, the provision of public access to speech corpora and the possibility to use them for different purposes, standardisation and creation of a computer set of tools for accumulation, processing and verification of speech databases [3]. To solve these problems, in the 1990-s, special coordinating centres were set up for recording, keeping, spreading and creation of public and standardised language recourses, including speech ones.

Among them there are:

- LDC (Linguistic Data Consortium, <http://www ldc upenn edu>)
- CSLU (Center for Spoken Language Understanding, Oregon Graduate Institute
- <http://www CSLU org edu>)
- ELRA (European Language Resources Association, <http://www elra info>)

The collection of speech corpora offered by the above centres is increasing every year, and more and more specialists are participating in their development. At the same time, there is a growth in capacity, diversity and computer equipment of the corpora themselves (further information on language resource centres can be found in review articles [4, 5]).

3. Speech corpora classification. Experience of creation and usage of speech corpora helps to identify a number of parameters that may be used as a basis for classification of speech databases and taken into consideration while designing a new corpus. The most important characteristics are (see also [4-6]):

- **purpose of a corpus usage:** specialised, general (representative), educational-illustrative;
- **type of speech material:** discrete speech, continuous reading speech, spontaneous speech, special dialogues;
- **type of text material:** lists of words/syllables, sets of separate sentences, coherent texts; monothematic or polithematic;
- **type of information associated with speech signal (annotations):** spelling, phonemic/phonetic transcription, prosodic transcription, acoustic-phonetic signal labeling: “events-based”, segment-based, prosodic, presence of other types of linguistic annotations and commentaries, for example, on speaker’s individual peculiarities of pronunciation or emotional content of speech fragments;
- **type of statistical provision** of sound language units: natural, even, representative, according to a special statistical scheme;
- **presence and type of auxiliary signal information** included into a corpus together with a speech signal: simple, multimodal and special corpora.

4. Russian speech corpora. As a rule, speech databases are monolingual. Speech corpora have been created not only for all technologically important languages (American English, German, Japanese, Chinese), but also for the majority of official languages of the European Union: British and Scottish variants of English, Dutch, Swedish, German, French, Italian, Spanish, there are also several multilingual corpora. As a result of a programme Copernicus ELRA offers also speech corpora for Eastern European languages (Polish, Bulgarian, Estonian, Romanian and Hungarian). The Internet site of the European Association also announced Russian speech corpora. As far as we know, a St. Petersburg company “Oditel” took part in their creation.

4.1. Speech corpus ISABASE. At the end of the 90-s, the Institute of System Analysis of the Russian Academy of Sciences (ISA RAN), together with specialists from a speech group of Philological Faculty of Moscow State University, created the first representative Russian speech corpus with labelling of speech fragments into sound units, which was used not only for research purposes, but also for the creation of automatic system of discrete speech recognition [6]. The corpus is monosignal, other parameters are shown below in Table 1.

Table 1. Parameters of Russian speech corpus **ISABASE**.

<i>Type of speech material</i>		Discrete speech	Dictors/speech fragments-sentences	Total size
<i>Text material</i>	<i>1</i>	Phonetically balanced set of 500 short sentences, monothematic	5 male dictors and 4 female dictors; 1863 fragments	4653 speech fragments; 3713 different words
	<i>2</i>	Phonetically representative set of sentences taken from literary texts; polithematic	15 male dictors and 14 female dictors; 3280 fragments	
<i>Annotation types</i>		Text of a speech fragment, phonetic transcription, the results of a manual segmentation of signal into words and phonemes	Transcription system of 110 monophones	

4.2. Speech corpus RuSpeech. In 2000-2001, ISA RAN, by order of Intel Corporation, also created so far the most representative Russian speech corpus, which can be used for the development of systems of Russian speech recognition [7]. General corpus characteristics are shown in Table 2.

Table 2. Parameters of Russian speech corpus **Ruspeech**.

<i>General characteristics</i>	<i>Type of speech material</i>	<i>Set of segments/sentences</i>	<i>Dictors/fragments</i>
	continuous speech; monosignal	50 hours of recording; 30 CD, more than 15 Gb; more than 50000 fragments-sentences	237 dictors: 127 men and 110 women; different age and education
<i>Text material</i>	<i>1</i>	Phonetically balanced set; polithematic	70 sentences, providing total (≥ 3 times) monophone covering;
	<i>2</i>	Phonetically representative (at allophone level) set of sentences taken from newspaper and news texts on Internet sites; polithematic	3060 sentences, providing a total coverage of allophones from a representative set;
		2000 phonetically diverse sentences;	20 dictors: 10 male and 10 female, 200 random sentences each; every sentence was pronounced by 14 dictors;
<i>Annotations</i>	Text of a speech fragment, standard and factual transcriptions, checked by experts; data on speaker and expert-phonetician	Transcription system of 114 monophones	

Besides the speech database itself, an important result of the **Ruspeech** project is an adjusted technology of speech corpora creation and a set of tools for its development [14, 15]. Among them, the following should be noted: the adjustment of Russian speech automatic transcriber; the creation of a program for the preparation of a text material with required phonetic and statistical characteristics; the

creation of an automated working- place of an expert-phonetician; batch program for speech recording; several programs for verification of the results of all development stages.

REFERENCES

1. Hunt A., Black A.W. Unit selection in a concatenative speech synthesis system using a large speech database // *ICASSP-96*, vol. 1, pp. 373-376, 1996.
2. Gibbon, D., Moore, R., Winski, R. (Editors). Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, 1997.
3. Krivnova O.F., Zakharov L.M., Strokin G.S. Speech Corpora (Experience of Development and Usage)// Proceedings of Dialogue'2001 Seminar on Computational Linguistics and its Applications. M., 2001 (in Russian).
4. Bogdanov D.S., Krivnova O.F., Podrabinovich A.Ya., Farsobina V.V. Database of Russian Speech Fragments ISABASE.// Intellectual Technologies of Information Input and Processing. M., Editorial URSS, 1998 (in Russian).
5. Bogdanov D.S., Bruhtiy A.V., Krivnova O.F., Podrabinovich A.Ya, Strokin G.S. Technologies of Speech Databases Development// Organizational Control and Artificial Intelligence. M., Editorial URSS, 2003 (in Russian).
6. Arlazarov V.L., Bogdanov D.S. Krivnova OI. F., Podrabinovitch A. Ya. . Creation of Russian Speech Databases: Design, Processing, Development Tools // International Conference SPECOM'2004. Proceedings. S-Pb. Russia, 2004. Pp: 650-656.