The Problem of Choice and Preparation of a Text Material for Speech Corpora

O.F. Krivnova

Philological Faculty, Lomonosov Moscow State University, Vorobyovi Gori, 1-st Building of the Humanities, Moscow, Russia, 119899
okri@philol.msu.ru

Abstract
Many tasks connected with speech analysis involve the problem of description, modelling and evaluation of the acoustical variability of sound units in various speech texts. Nowadays, a large-scale and statistically representative research of the acoustic variability of sound units has become possible thanks to quickly developing computer technologies and the use of representative speech corpora, containing special annotations. This paper is dedicated to the principles of the selection of a text material for scientific and applied modelling of acoustic variability. Computer tools necessary for the creation of a statistically representative and phonetically reliable speech corpus are also discussed; the author’s experience of a text material preparation for Russian speech corpora is briefly described.

1. Introduction
Experimental analysis and modelling of acoustic variability of sound units in natural speech is of considerable interest for both general phonetics and applied projects in the field of speech technologies. In Russian phonetics, despite a rather significant amount of researches on this problem, one cannot say that it is completely solved even for such relatively simple speech conditions as reading of common phrases and small coherent texts. The growth of empirical basis of phonetic research and the inclusion of various speech material in it motivate scientists to seek new directions and possibilities to model the acoustic picture of natural speech. There is a constantly increasing interest in the development of statistically representative speech corpora, equipped with special linguistic and other annotations (speech databases). It is known that the creation of annotated and structured speech corpora was initiated in the 1970s by works on automatic speech recognition. The quality of modern statistical recognition algorithms directly depends on the capacity of such corpora [1]. It is essential for researchers with so-called engineering approach, based on the study of speech signal fine structure, to test their hypotheses on a large quantity of representative annotated data. A modern approach to a text-to-speech synthesis, which uses concatenation method, also presupposes the use of large speech corpora [2]. According to the experts [3], a corpus-based approach is crucial for the development of the technologies of synthesis, especially for modelling prosodic parameters and peculiarities of a speaker’s voice.

2. Preparation of a text material for a speech corpus
The creation of speech databases is a technologically complicated process, which consists of several main stages [4]. Some tasks require phonetical expert knowledge. Among these tasks are:
• the preparation of phonetical basis for forming a speech corpus;
• the development of standards for speech transcription at different levels;
• the preparation of a text material for obtaining necessary sound signals.
The present paper is dedicated to the latter task. As far as the former two tasks are concerned, they have already been discussed in the author’s works [5, 6]. It can be reminded that in the process of preparation of a phonetical basis, it is necessary to choose a transcription system in order to create a standard (expected pronunciation) phonetic transcription for corpus sentences. To work out large-scale databases, it is essential to have and use a reliable automatic transcriber for text transcription. During the preparation of a text material such program allows to estimate in advance expected phonetical characteristics, including statistical ones, of a future speech corpus.

2.1. The sources of a text material
In applied projects, a text material for a speech database usually consists of a number of independent sentences, corresponding to sound files, included in a corpus. Such aggregate is prepared in advance and may consist of several sets of sentences, organized differently depending on concrete criteria applied to their phonetic characteristics. Some sets may even represent coherent texts. Here we are not concerned with the problems of creation of spontaneous speech corpora – it is a different and complicated topic.

Sets of sentences forming a corpus text material can be obtained in various ways, depending on particular spheres and purposes of the corpus’ application. Sentences may be selected or constructed by linguists specially for a corpus, like in famous American corpus TIMIT, where sets of sentences satisfy certain phonetical conditions, chosen beforehand. Thus, the first two sentences in this database are created to reflect as much as possible the dialect features of speakers’ pronunciation. Sentences can also be taken from newspapers, magazines and Internet news sites. The articles’ subjects may vary or, on the contrary, be limited. Sources of this kind usually require a
certain correction of length and grammatical structure of original sentences.

Finally, a corpus may consist of rather extensive abstracts of a coherent literary text, including dialogs, e.g. taken from plays.

2.2 Phonetical composition of a text material

Demands made to a phonetic composition of a text material can vary and are specified by particular aims, for which a speech database is developed.

Let us consider some of the most popular requirements of this kind. One of them may demand a text material to meet some of the criteria of phonetical representativeness. For example, it can require the transcribed representation of a corpus text to contain all the phonemes of a particular language (monophones) and all their main (primary) allophones. Besides, the frequency of occurrence of each monophone could be claimed to be not less than a certain, estimated beforehand number. There may be other requirements concerning phonetical representativeness, for instance, the availability of the combinatorial allophones of phonemes (i.e. monophones differing in their right and left segment contexts, so-called thriphones).

To meet the latter requirement, one can, of course, endlessly increase the number of sentences in the text material. However, a corpus usually has certain limitations on the amount of sentences in it. To satisfy these requirements, it is possible to apply special automatic iterative procedures to filter the chosen sentences and texts, which enable to obtain the optimal realization of any scheme of statistical balancing of corpus phonetic elements.

2.3 Methods and tools of the optimal selection of texts

The task of the optimal selection of texts usually consists in the creation of a minimal or limited text set, satisfying certain apriori statistical requirements of the frequency of occurrence of particular elements in a language and/or corpus (e.g., phonemes, monophones, thriphones in the case of a speech corpus). To fulfill this task, one can use statistical procedures, based on methods briefly described below.

2.3.1. Accumulative method

The so-called accumulative method of a text corpus creation consists in the following.

A list of units (in our case – sound units) in an already formed text corpus is made. In the beginning this list is empty. Besides, there is a number of candidate sentences to be included in the corpus and provided with transcriptions (initial text pool). In the process of analysis of each sentence in the pool, a list of a sentence’s sound units is created. Then it is compared with a list of units in the already included in the corpus sentences, and a number of “new sounds” that the sentence, if added, contributes to the general list (its phonetical contribution) is estimated. A sentence is included in the corpus only if its phonetical contribution to it is not less than a certain number. The last is a filtration parameter, set before the process of selection. At the initial stage this number can be quite big (e.g., 10). If the number of sentences chosen at this stage of selection is less than necessary, the level of filtration is lowered and rejected candidate sentences are analysed again.

The result of the application of the accumulative method strongly depends on the order of sentences in the pool. It is a considerable disadvantage of this method. To optimise the procedure of selection, a simple accumulation can be replaced with a more perfect method, known as a “greedy” algorithm [7].

2.3.2. “Greedy” algorithm

To discuss the essence of this method, let us take as an example the creation of a rather small set of sentences, where every sound from some big initial pool of texts is represented at least once.

First, for each sentence in the initial text pool it is generated a list of its sound units, with repeated elements being removed. Then a so-called “greedy” algorithm for the selection of required sentences is applied. A dictionary on computer systems defines a greedy method as “an algorithm of completing a certain task where each stage does its best to approach the achievement of the task’s aim as quickly as possible”. In other words, in this method possible future benefits are sacrificed for the sake of immediate goal achievement.

A “greedy” algorithm looks for a sentence in the initial text pool that has the longest “sound” list. After that all the elements in this sentence’s list are removed from the other sentences’ sound lists, and the sentence itself is added to the future corpus (in the beginning this corpus is empty). “Sound” lists of the remaining sentences now contain only units, absent in the future corpus. Among them the next sentence with the longest “sound” list is chosen, and so on, see above. The described procedure is repeated as long as all sound units are removed from the the remaining sentences’ lists, i.e. until all these lists are empty. As a result, the corpus of selected sentences will be close to the minimal sentence set with all sound units from an initial text pool included.

There exist various modifications of the described algorithm. For instance, it may happen that a full coverage of all sound units in the initial text pool will require a too large corpus. In this case, there may arise a task to create a corpus of a certain size, which contains only most frequently occurring units. To achieve this aim, it is possible to use a weighed “greedy” algorithm. In a regular greedy algorithm the choice of the next sentence is based on the calculation of the number of various units in it, necessary for the corpus. In a weighed greedy algorithm, each sound unit has a certain weight coefficient, and the decision on choosing the next sentence is based on the calculation of the sum of all sound unit weight coefficients in this sentence.

The application of a weighed modification of the algorithm to forming a corpus with the full coverage of all sounds from an initial pool can help to reduce the size of the corpus.

3. Representative speech corpora for Russian

A statistical software, based on the discussed principles, was worked out by the speech group of philological faculty of Moscow State University [6,8] and was used in the development of a large-scale Russian speech corpus Ruspeech in cooperation with the Institute for System Analysis RAS and Cognitive Technologies, Ltd. for Russian speech recognition project of Intel Corporation [4]. This software tools was also applied in a research project on Russian phonetics [9].
So far the speech corpus RuSpeech developed in 2000-2001 is the most representative Russian speech corpus. It contains more than 50 hours of recorded and transcribed continuous Russian speech.

Transcription code system of RuSpeech includes 114 monophones and is constructed on the basis of the popular Russian transcription code system offered by R.I.Avanesov [10]. One of the main requirements to the corpus was to provide full phoneme coverage per speaker and full allophone coverage integrally in main sets of the corpus. Besides that the corpus represent theoretically natural distribution of phoneme-monophones by which we meant the frequencies of language units statistically defined on a large set of samples.

The text material used in Ruspeech database consists of four following separate sets of sentences:

A set of 70 sentences which provides full phoneme coverage. Each phoneme is represented by at least 3 samples. The sentences of this set were pronounced by each speaker in the train set of the corpus. The set was constructed by the linguists to provide its phonetic completeness.

A set of 3060 sentences was also designed to train the speech recognition software. It provides full allophone coverage (representation of each allophone more a twice). These sentences were even distributed among speakers in the train set.

Two sets (each contained 1000 sentences) were designed for development of speech recognition software components and for testing of the entire system.

The last three sets of sentences were made up of texts from different Russian newspapers or Internet news sites. The stories were selected so that they would cover various subjects in politics, economy, culture, art, medicine, sports, etc. Some sentences were taken as they occurred in the text, while others had to be slightly changed to be as a rule no longer than 9–10 words. The sentences were selected with the automatic filtering procedure described above to meet the requirement of sufficient representation of allophones. Since there were no special selection of words these sets also meet the requirement of theoretically natural phoneme distribution.

4. References


