Compiling of Phonetic Database Structure

Maya Heydarova [1]

[1] *Azerbaijan National Academy of Science İnstitute of Linguistics named after Nasimi*
31 H. Cavid, Baku, AZ1143, Azerbaijan

DOI: 10.22178/pos.69-6

LCC Subject Category:
PE1001-1693

Received 16.03.2021
Accepted 26.04.2021
Published online 30.04.2021

Corresponding Author:
mayahaciyeva.mh@gmail.com

Abstract. The voice corpus of language is the essential part of the linguistic resources, and it contains the phonetic database. A phonetic database is a structured collection of software-delivered speech fragments. Nowadays, phonetic database or voice corpus became like a new element in speech technologies, and much investigation has taken place according to this event. The investigators' interest in voice corpus is related to the development of a speech recognition system. Today it is enough to experience in preparation of a phonetic database. Equipped with unique information on the preparation and usage of everyday speech corpus, the development level of speech technologies and the increasing power of computer technologies allow for the investigation of various language materials, largescale, and statistical phonetic research. These developed directions of linguistics were investigated in this article. Speech corpora are a valuable source of information for phonological research and the study of sound patterns. The study of speech corpora is in its infancy compared to other field studies in linguistics. Existing speech corpora form the part of the world's languages and do not fully represent all the dialects and speech forms by phonological aspect. The article analyses the history, structure, and importance of developing speech corpses, a branch of corpus linguistics and has developed in recent years. The article also lists the main features to be considered in the design of the speech corpus.

Keywords: phonetics, sound corpus, database, speech technology, speech signal.

INTRODUCTION

In the modern world, the penetration of information technology in all spheres of human activity coincides with the penetration of globalisation into all processes taking place in the world. The interdependence of modern society with information technology has become a key feature of human relations, including exchanging information. In this regard, information technology is actively used in various fields of science in the world.

One of such areas of activity is language corpus and linguistic databases prepared by linguists in applied linguistics.

Since the sound corpus is a phonetic database, the topic relevance occupies a special place in the linguistic database and is its main component. Phonetic database research is one of the minor studied issues in linguistics.

The primary purpose is to give an accurate classification of the principles of phonetic database structure and conduct analysis regarding theoretical information within the article.

Since the sound corpus of many languages, including Azerbaijani, is still in its infancy, studying the experience of world languages can provide researchers with new materials.

In addition to mathematical and statistical methods, which are modern research methods, the article used the descriptive method. The study refers to the scientific research of authors who have researched this topic before us in this area.

A linguistic database is a linguistic information resource, and in addition to public databases, it identifies different levels of language. The term "linguistic resources" was first used by the Italian scientist A. Zampolli in 1992 [12]. The term was chosen to express an opinion about the linguistic information and descriptions used for creating and developing compelling text and word processing systems.

Linguistic information resources are an integral part of information resources. The information

resources are passive (books, newspapers, dictionaries, encyclopedias, databases and banks, etc.) and active (algorithms, models, programs, knowledge bases) [9].

Language materials posted on the Internet are called linguistic information resources. The linguistic resources are linguistic databases that can be updated automatically (entering new data, deleting or changing old data) and searching for this information. The linguistic database, an integral part of linguistic information resources, consolidates information about world languages into one database.

The database (DB) is a specially structured place used to store various data types (text, numbers, time expressions, currencies, etc.) in the computer's permanent memory.

The linguistic resources are essential for personal computers (PC) users and computer systems involved in text and word processing. In particular, they are involved in recognising speech and anonymous texts, abstracting, annotating and translating texts, constructing dialogue texts, automatic analysis of text, synthesis of text and speech, and so on [8].

Linguistic databases are divided into phonetic, lexicographic and grammatical levels.

The phonetic database was first funded by the US Department of Defense in the 1980s. With the support of the US Department of Defense, the TI-DIGITS corpus was developed in 1984 to test isolated number and number sequence recognition systems. Later, phonetic data corporations, such as the King Corpus, were developed to analyse Road Rally keywords and identify the speaker. TIMIT, which serves as a prototype for other voice corporations developed under the US Department of Defense's State Program for the Development of Linguistic Technologies, the Resource Management and Wall Street Journal for speech recognition research, as well as dialogue projects such as the Air Travel Information Service have also been included in the systems to understand natural language and study spontaneous speech [12].

From this, we can conclude that to make sound corpora, it is necessary to overcome complicated technological issues. This requires significant financial support, staff training, standardisation, and computer tools to ensure the accessibility and versatility of speech corpora and the collection, processing, and verification of speech databases. To solve these problems, special centres were set up in the 1990s to collect, store and distribute the standardised language and speech resources with public access.

LDC (Linguistic Data Consortium) [4], CSLU (Center for Spoken Language Understanding, Oregon Graduate Institute) [3], ELRA (European Language Resources Association [2]), and other databases can be used as examples.

In recent years, phonetic databases or the sound corpus have been identified as a new element in speech technology. Researchers' interest in creating a sound corpus arose when confronted with the acoustic variability of the sound units of a language with different sources; thus, beginning the speech recognition field. Modern recognition systems are studied based on many speech materials, consisting of audio recordings of many speakers. The sound corpus is needed to solve many scientific problems related to the analysis and description of speech in different languages.

"For most announcers, the need to develop more accurate models of speech flow elements that maintain their quality has led to the creation of specialised speech VBs that include the number of announcers, material according to different pronunciation types, system vocabulary and frequency of each element (number of manifestations)" [6].

Due to the small number of scientific and theoretical databases on the subject, the research needs to be based more on experience. However, in recent years, some theoretical work has been done in this area. In particular, it is worth noting the dissertation [6]. In addition, you can find interesting information about this field in the literature and articles were written about the public database [11, 13].

RESULTS AND DISCUSSION

The primary purpose of the phonetic database is to search and test hypotheses about the characteristics of speech sounds, check and adjust the parameters of models and methods for speech signal analysis, and establish systems for automatic speech recognition and synthesis. The phonetic database is used to study the phonetic features of speech at the segmental and suprasegmental levels, develop systems for automatic recognition and synthesis of speech, and check and identify a person's voice biometric parameters.

"The emergence of technology for the technical modelling of the speech signal by the process of covert marking, which involves the stochastic nature of the signal, led to the creation of a special database. The main purpose of phonetic databases is to provide information on the distribution of parameters of sound unit models, to define and adjust the parameters, as well as to check the operation of recognition systems (assessment of accuracy)" [10].

The phonetic database includes:

1. speech materials, all possible variants of sound units that maintain the natural frequency of occurrence;

2. to describe this material in transcribed and marked (marked) orthographic writing and connect each writing unit with the corresponding part of the acoustic signal.

To develop a speech recognition and identification system based on the speaker's voice and speech. It is necessary to analyse and summarise a large amount of information about speech signals. Such information can be obtained from large phonetic databases.

4.1. Speech database and voice recognition system

The term "speech database" is usually used to describe the improvement and evaluation of systems and algorithms in linguistic technology applications for processing speech and language materials and identifying a large set of linguistic data and descriptions provided in electronic form [8].

An example of such an electronic array is an electronic encyclopedia for phonoscopists.

The database of the electronic encyclopedia prepared by R. Potapova contains the following information:

– theoretical bases of language and speech (natural language, literary language, local dialects, sociolects, jargons, slangs, types of people language, bilingualism, language interference, national language, speech communication, types of speech, spontaneous speech, speech culture);

– basics of speech formation mechanism (anatomical-physiological, mental, intellectual, linguistic and extralinguistic speech formation mechanisms, initial and acquired speech habits, phonation, specificity of sound formation and sound quality, articulation and coarticulation, segment and suprasegment level);

– basic concepts of linguistic, paralinguistic and extralinguistic speech information (phonetic-phonological, lexical, syntactic, semantic, pragmatic and phonostylistic levels of speech consideration, defects of the speech tract, those that lead to voice changes, psychological state and neurophysiological features of speech);

– description of algorithms for mathematical processing of speech signals, calculation and comparison of acoustic signs, the essence of the features of inter and intra-speaker variability, statistical decision-making rules and their adaptation to relearning;

– arrangement of linguistic identification features of the speaker's oral speech, their selection and comparison methods;

– instructions for the expert - phonoscopist conducting identification research (computer input and segmentation of speech signals, assessment of phonogram quality, selection and comparison of acoustic and linguistic features of speech flow, different expressions, words and sounds, decision making, compilation of expert conclusions); typical examples of expert (phonoscopist) results, expert terminological dictionary [8].

There is a need to develop broad, informative speech corpora for phonetic research and computer tools and applications for their development and ease of use. Speech recognition systems with the highest reliability and performance are mainly based on statistical modelling methods of speech and language events. This requires a voice recording of at least 100 announcers. Thus, the preparation of phonetic databases consists of the following stages:

a) Development of an application interface for working with databases. Writing an application based on the MS Visual Basic programming language to work with the database (key elements - add, delete, search the database under certain conditions (by gender, nationality, age, etc.).

b) Preparation of the database project. Development of this project based on MySQL (file folder, essential operations and transactions) database.

c) In the next step, add new elements to the database and add additional parameters to the database by search (by sounds, syllables, etc.).

d) Develop a structure for linguistic interpretations, including transcription of speech patterns.

The preparation of the program part consists of the following stages:

1. Sound recording – sounding of fragments of speeches of different lengths;

2. Storage of personal information about announcers;

3. Storage of text documents as an analogue of sound files;

4. Division of speech samples into phrases, syntagms and words, transcription and storage of the results of expert phonetic analysis;

5. Preparation of phonetic dictionary (according to phoneme, syllable, rhythmic accent);

6. Search for speech information by various features (by name, gender, age of the announcer; by a regional variant of language (according to dialect); by keywords, sound, syllable) [13].

4.2. The preparation and use of speech or sound corpora

The experience gained in the development and use of speech corpora allows us to identify some features that can be the basis for the classification of speech databases and must be considered when designing a new corpus. Let's look at the most important of these features:

– type of speech material: discrete speech, continuous reading speech, spontaneous speech, particular dialogues;

– type of text material: list of words or syllables, individual sentence sets, related texts; monothematic or polythematic;

– type of speech signal: laboratory speech, office speech, public speech, telephone speech (ordinary or mobile phone; radio, tele-speech);

– type of information related to the speech signal: spelling, phoneme or phonetic transcription, prosodic transcription;

– acoustic-phonetic marking of the signal: the presence of segment, prosodic, other linguistic summaries and interpretations;

– type of statistical balancing of sound units of language: according to a particular statistical scheme natural, equal;

– availability and type of additional signal information entered into the corpus by a speech signal: simple multimodal and particular case.

The work involved in interpreting speech is extremely laborious, but it can be done reasonably because of special programs.

The following software tools are used during the preparation of the sound corpus [1]:

1. Praat professional phonetic analysis program

2. Professional audio editor Sound Forge 8.0

3. ELAN multi-level linguistic annotation program

4. Lexicographic data processing program KartaTeka

5. Access, SQL, MySql, etc. programs for database preparation.

The development of sound corpus and their multilevel marking has recently become possible due to the development of information technology in the humanities.

In recent years, Azerbaijan has accumulated some practical experience in speech recognition, especially during forensic phonoscopic examination, and some scientific studies have been conducted in this regard [7, 8].

CONCLUSION

At present, enough experience has been gained to develop a phonetic database. The ability to store large amounts of digital information on modern computers and high-quality data processing, high quality digital audio and acoustic processing of digital audio materials, advances in the development of databases and full-text search engines allow to create of more sophisticated phonetic databases. The development and use of everyday speech corpora equipped with specialised information, the development of speech technology, and computer technology's growing power will allow researchers to conduct large-scale and statistical research on various speech materials. Especially if we consider the lack of research in this field in many newly independent countries, including Azerbaijan, we will see how much research in this area and their results are needed in linguistics. In this area, the study of databases of advanced languages and the study results are significant for future research.

REFERENCES

1. Bogdanova, N., Asinovskij, A., Rusakova, M., Ryko, A., Stepanova, S., & Sherstinova, T. (2009). Zvukovoj korpus kak sposob monitoringa i fiksacii raznyh form estestvennogo jazyka [Speech corpus as a tool for monitoring and fixation of various forms of natural language]. Retrieved from http://www.dialog-21.ru/digests/dialog2009/materials/html/07.htm (in Russian) [Богданова, Н., Асиновский, А., Русакова, М., Рыко, А., Степанова, С., & Шерстинова, Т. (2009). Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка. URL: http://www.dialog-21.ru/digests/dialog2009/materials/html/07.htm].
2. ELRA. (2021). *About*. Retrieved from http://www.elra.info
3. Heeman, P. (2016). *CS550 Spoken Dialogue Systems*. Retrieved from https://cslu.ohsu.edu/~heeman/cs550/
4. Krivnova, O., Zakharova, L., & Strokin, G. (2001), Rechevyje korpusy (opyt razrabotki i ispolzovanie) *Dialog*. Retrieved from http://www.dialog-21.ru/digest/2001/articles/krivnova (in Russian) [Кривнова, О., Захаров, Л., Строкин, Г. (2001). Речевые корпусы (опыт разработки и использования). *Диалог*. URL: http://www.dialog-21.ru/digest/2001/articles/krivnova].
5. Linguistic Data Consortium. (2021). *About*. Retrieved from http://www.ldc.upenn.edu
6. Loseva, E. (2006). *Formirovanie mnogojazychnoj foneticheskoj bazy dannyh (primenitel'no k rechevoj realizacii vibrantov)* [Formation of the multilingual phonetic database (in relation to speech realization of vibrant)] (Doctoral dissertation). Moscow: n. d. (in Russian) [Лосева, Е. (2006). *Формирование многоязычной фонетической базы данных (применительно к речевой реализации вибрантов)* (Кандидатская диссертация). Москва: n. d.].
7. Musayev, H., Əliyev, L., & Vəliyev, H. (2020). *Audio və videoyazıların məhkəmə kriminalistik ekspertizası (elmi-praktik vəsait)* [Forensic medical examination audio-and videos (scientific and practical grant)]. Baki: Məhkəmə Ekspertizası Mərkəzi (in Azerbaijanian).
8. N. d. (2021). *Ponjatie foneticheskoj bazy dannyh. Trebovanija k sovremennym foneticheskim bazam dannyh dlja fundamental'nyh i prikladnyh issledovanij* [Concept of the phonetic database. Requirements to modern phonetic databases for basic and applied researches]. URL: https://studexpo.ru/758977/literatura/ponyatie_foneticheskoy_bazy_dannyh_trebovaniya_sovremennym_foneticheskim_bazam_dannyh_fundamentalnyh_prikladnyh (in Russian) [N. d. (2021). *Понятие фонетической базы данных. Требования к современным фонетическим базам данных для фундаментальных и прикладных исследований*. URL: https://studexpo.ru/758977/literatura/ponyatie_foneticheskoy_bazy_dannyh_trebovaniya_sovremennym_foneticheskim_bazam_dannyh_fundamentalnyh_prikladnyh].
9. Ostrejkovski, V. (2000), *Informatika* [Informatics]. Moscow (in Russian) [Острейковский, В. (2000). *Информатика*. Москва: Высшая школа].
10. Potapova, R. (1997). *Rech: kommunikatsija, informatsija, kibernetika* [Speech: communication, information, cybernetics]. Moscow: Radio I Svjaz' (in Russian) [Потапова, Р. (1997). *Речь: коммуникация, информация, кибернетика*. Москва: Радио и связь].
11. Potapova, R., & Potapov, V. (2018). Rechevye bazy dannykh kak chast' multimodal'nykh korpusov [Spoken language databases as a part of multimodal corps on the Internet]. *Vestnykh MGLU. Gumanitarnye nauki, 6*(797), 99-116 (in Russian) [Потапова, Р., Потапов, В. (2018). Речевые базы данных как часть мультимодальных корпусов в Интернете. *Вестник МГЛУ. Гуманитарные науки, 6*(797), 99–116].
12. Zampolii, A. (1998). Introduction of the General Chairman. In *First International Conference on Language Resources & Evolution*, 28-30 May (pp. 15-25). Granada.
13. Zav'jalova, V. (2010). Znachimost' spetsializirovannykh rechevykh baz dannykh dlja formirovanija foneticheskoj kompetentsii [Application of specialized speech databases for

developing phonetic competence]. *Vestnik Irkutskogo gosudarstvennogo lingvisticheskogo universiteta, 3*, 151-156 (in Russian)
[Завьялова, В. (2010). Значимость специализированных речевых баз данных для формирования фонетической компетенции. *Вестник Иркутского государственного лингвистического университета, 3*, 151–156].