Journal of Computer Science

Review

Speech Corpora for Different Languages: A Systematic Review

Vladimir Igorevich Fedoseev, Anton Aleksandrovich Konev and Natalia Sergeevna Repyuk

Department of Complex Information Security of Computer Systems, Tomsk State University of Control Systems and Radio Electronics, Tomsk, Russia

Article history
Received: 29-07-2024
Revised: 30-06-2025
Accepted: 17-07-2025

Corresponding Author:
Natalia Sergeevna Repyuk
Department of Complex Information Security of Computer Systems, Tomsk State University of Control Systems and Radio Electronics, Tomsk, Russia
Email: rns@fb.tusur.ru

Abstract: The study of speech signals relies on carefully curated audio recordings, which are compiled and stored within specialized speech corpora. This article provides a comprehensive overview of such corpora across multiple languages, with particular focus on Russian, English, and Arabic. It notes that Russian and Arabic are represented by fewer corpora compared to the more extensive resources available for English. The discussion includes an examination of typical speech corpus structures, a description of standard parameters for characterizing corpora, and an outline of common metrics used to describe the speech signal itself.

Keywords: Dataset, Pronunciation, Speech Corpora, Transcript, Speech Recognition

Introduction

Nowadays, people everywhere use speech technology. Speech technologies include the perception of the meaning of phrases, speech imitation, speech-to-text conversion, and speaker identification by voice. Speech technologies are difficult to learn because they combine different disciplines: Computer science, mathematics, programming, and linguistics. To create and improve speech technologies, researchers study the parameters and features of the speech signal. Speech corpora are used to study and analyze speech signals. The speech corpora is a database with audio recordings that are brought to a certain structure.

The purpose of writing this article is to study the structure of different speech corpora for different languages. To achieve this goal, existing speech corpora will be studied, taking into account their structure, and the following research tasks will be solved.

RQ 1: What are the typical parameters used to describe speech signals? The purpose of this study is to determine the features characterizing the speech signal.

RQ 2: What are the typical parameters used to describe speech corpora? The goal is to understand which characteristics can be used to describe the speech corpus.

Research Questions

In this review, the main attention was paid to speech corpora in Russian, English and Arabic, as the further research activities of the authors are related to them. The number of articles on the keyword “speech corpora” amounted to 2,844 articles in various journals that have been published in the Mendeley database over the past 3 years. Of these, 233 articles were found about Russian corpora, 617 about English corpora and 93 about Arabic corpora. Figure 1 shows a diagram showing the quantitative ratio of the sources found.

In this review, 69 sources were reviewed, 20 of them on Russian corpora, 28 on English corpora, 7 on Arabic corpora and 14 on others. Figure 2 shows a diagram showing the quantitative ratio of the analyzed sources.

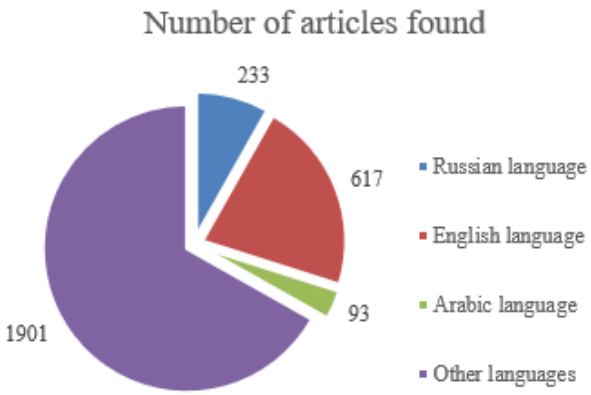

Fig. 1: Quantitative ratio of the found publications

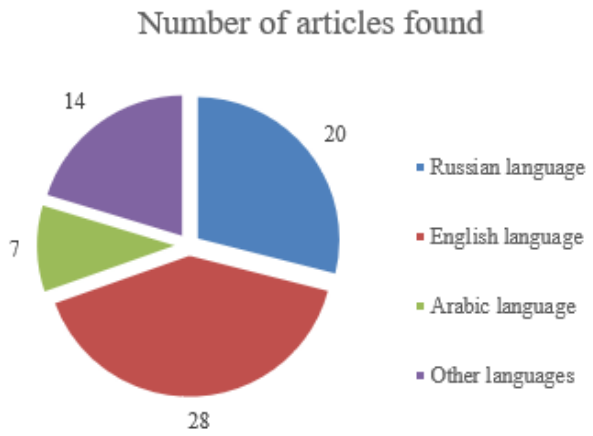

Fig. 2: Quantitative ratio of the analyzed publications

Russian Speech Corpora

Saito *et al*. (2021) used and described the Mozilla Common Voice Dataset corpus. This dataset was developed for any speech handling systems, including recognition. For English in version 11.0, the dataset contains 2320 verified hours of audio material (i.e., audio material with all metadata) recorded by 84673 speakers of both genders. The recordings of this dataset are noisy, to further test the effectiveness of the recognition algorithm. For Russian, this corpus contains 215 recorded and 180 verified hours of audio material recorded by 2731 speakers. The dataset was developed for machine learning systems, so the recordings are divided into three samples: Training, refining and testing. There is also a subdivision into tested, untested and "other" records.

Park and Mulk (2019) have developed a corpus for speech synthesis systems. Audio books were used as a data source. The dataset was created for ten languages: Dutch, German, Finnish, French, Hungarian, Spanish, Russian, Greek, Spanish, and Japanese. For each language, recordings of only one male speaker were used. Phrase lengths ranged from 5-11 seconds.

The Russian corpus, presented by Tatarinova and Prozorov (2018), is designed for speech recognition systems. Radio program recordings were used as the data source. A total of 132 recording Training Speech Enhancement Systems with Noisy Speech Datasets were selected. They contain from 17 to 141 words. The total volume of audio material is 1 hour 12 minutes 29 seconds. It is noted that this speech corpus is at an early stage of development. Also mentioned in the article are the following datasets: RuSpeech, Dream Stories, and TIMIT Acoustic-Phonetic Continuous Speech Corpus.

Ru Speech was developed by Arlazarov *et al*. (2004) for speech recognition systems with an emphasis on noise elimination. The number of speakers who participated in the recording was 203 speakers, 111 of them men and 92 women. Each speaker utters 250 sentences, 70 of which are the same for all and cover all phonemes, and the rest are selected from a set of sentences so as to cover all allophones. Phoneme-by-phoneme segmentation is present. The total amount of audio material is 50 hours. Factors such as gender, age, accent, place of birth and residence were taken into account in the recording. The dataset is designed for machine learning systems and consists of four samples: Train, development, test and peculiarity. The last one contains atypical recordings of speakers with pronunciation problems.

Kibrik *et al*. (2009) developed “Dream Stories” for speech recognition systems. The corpus consists of 129 stories from children and adolescents from 7 to 17 years old about what they saw in their dreams. The stories were recorded immediately after waking up. The total duration of the audio is about 2 hours; the corpus volume is about 14,000 word occurrences. The stories are categorized into two groups: 60 stories were taken from children and adolescents from the control group, and 69 stories were taken from experimental participants with one or another neurotic disorder. The database consists of two tables: "List of stories" and "transcription" (full, simplified and minimal transcription can be chosen), which are connected with each other by an external key "code number". The "list of stories" table contains the fields "gender", "age" and "diagnosis" in addition to the primary key "code number". The table "transcription" in all variants has a double primary key "code number"-"number" ("number" of the record in the transcription). In addition, the table contains the fields "time" (duration of the fragment), "EDE" (transcription of the transcription) and "comment". There is no segmentation into individual phonemes.

Prodeus (2013) has developed the VitalVoice speech corpus. However, after searching for information about this corpus, it turned out that it is a software product of "Speech Technology Center" STC and it is outdated. From the documentation on this product at the moment it was possible to find only a user manual describing the process of user interaction with the graphical interface and does not contain any information about the corpus used, except that it was used for speech synthesis.

The National Corpus of the Russian Language (2004) is a large corpus of mainly texts. Its multimodal module is essentially a separate speech corpus with audio and video materials. The recordings used are fragments of films, oral public and non-public speech, authors and fiction reading, and theatrical speech. Collectively, the corpus has 187 thousand such recordings (as of 2015). The corpus takes into account many speaker parameters and other additional information: Gender, age, and data on the fragment's origin, sphere of functioning, topic, text type, and target audience. There is also a very detailed system for transcribing spoken words. In addition, if the excerpt used is a video, all gestures

occurring in it are also described. The speech corpus is used for language research and is not intended for speech recognition purposes, but contains a substantial amount of audio and video material with detailed descriptions and metadata.

Datatang (2022) Hours - Russian Speech Data by Mobile Phone was developed for speech recognition systems. Voice commands to a smart home or voice assistant spoken by a person into a cell phone were used as the data source. There were 1,960 speakers, half of whom were men and the other half were women; the 18-25 age group made up 61% of all speakers, the 26-45 age group made up 35 and 4% of the speakers belonged to the 46-60 age group. The total amount of audio material rendered in the title is 1002 hours.

Karpov *et al*. (2021) developed a Golos dataset in Russian for speech recognition systems. The speech corpus consists of recorded audio files, annotated manually on a crowdsourcing platform. The total duration of the audio is about 1240 hours. Since the dataset is designed for machine learning-enabled systems, its structure consists of record-annotation pairs collected in two samples: training and test. The samples are categorized into two domains "Crowd" and "Far field". The distribution of audio materials by recordings and hours is shown in Table 1.

Cattoni *et al*. (2021) developed Must-C for automatic speech recognition and translation systems. The corpus consists of 8 sub-cases for Spanish, German, Italian, Portuguese, Romanian, French, Dutch and Russian. For the latter, the volume of audio material is 489 hours and contains 270 thousand recordings with 1724 male and 859 female speakers respectively. The corpus materials used are recordings of TED Talks conferences translated into different languages. The structure of the recordings is an audio-transcript pair, where the transcript is in turn divided into the original English line that has been translated and the same line in the desired language. Word-level segmentation is present.

Wang *et al*. (2020) developed CoVoST for speech recognition and translation systems. It contains 15 subcases, including those for Russian and Arabic.

The recordings are audio-decoding pairs in two categories: English translation and English translation. Since the dataset was developed for systems with machine learning, all recordings are divided into three samples: Training, refinement and test.

Boito *et al*. (2019) developed MaSS for speech recognition and translation systems. It contains 8 languages: Basque, English, Finnish, French, Hungarian, Romanian, Spanish and Russian.

Table 1: Structure of Golos dataset

Domain	Train files	Train hours	Test files	Test hours
Crown	979 796	1 095	9 994	11.2
Farfield	124 003	132.4	1 916	1.4
Total	1 103 799	1 227.4	11 910	12.6

The recordings are an audio-decoding pair. The transcription includes a direct description of the spoken text as well as a translation into the remaining 7 languages. The materials used are recordings of Bible readings in each language. The number of speakers and their gender is not specified. The amount of audio material is also not specified.

CrowdSpeech is an English-language dataset that was developed for speech recognition systems. In parallel, another similar dataset VoxDIY for Russian language was developed (Pavlichenko *et al*., 2021). Recordings from the LibreSpeech and RusNews speech corpora were taken as source data, the text versions of which were read out by new speakers recruited using crowdsourcing. Since the corpus was designed for model training, its structure follows the train-dev-test model, but no such division is present for VoxDIY. An approximate partitioning by number of recordings, speakers and hours of audio material is presented in Table 2.

The descriptions of all the recordings in this speech corpus are stored in json format without segmentation. It represents a list of key phrases in transcription and string representation, a list of speakers pronouncing this phrase and a file containing the record.

Panayotov *et al*. (2015) developed the LibriSpeech dataset in English for speech recognition systems. The main feature of this dataset is that there are extended versions for other languages, including Russian. The corpus is designed for training the Kaldi speech model. The corpus itself is divided into three samples: Training, refinement and test. The recordings are short excerpts from audiobooks (no more than 35 seconds) accompanied by text transcripts. Public domain audiobooks are used as materials. The number of speakers, their gender, and the amount of audio material for each sample are shown in Table 3.

The above described speech corpora were analyzed with respect to the parameters under study.

Organized Data on Russian Language Corpora

All the considered corpus for the Russian language were analyzed in terms of the selected parameters: Scope, number of speakers, segmentation, speaker's gender and age, the presence of features.

Table 4 shows the collected data which were further processed. Some corpora have several variations. The table presents only those of them, the records in which are in Russian.

Segmentation is used in most cases, but it is often not phonemic but affects larger parts of speech. When creating a speech corpus, the gender factor is taken into account more often than the age factor. The speech characteristics of the speakers in these cases are not taken into account. Physical and emotional states are also not taken into account.

Table 2: Structure of datasets CrowdSpeech and VoxDIY

Dataset	Version	Mean Sentence Length, words		#Recordings	#Workers	#Answers
		Ground Truth	Crowdsourced			
CrowdSpeech	train-clean	34.6	32.8	11, 000	2, 166	77, 000
	dev-clean	20.1	19.5	2, 703	748	18, 921
	dev-other	17.8	16.8	2, 864	1, 353	20, 048
	test-clean	20.1	19.2	2, 620	769	18, 340
	test-other	17.8	16.8	2, 939	1, 441	20, 573
VoxDIY	RU	13.8	13.6	3, 091	457	21, 637

Table 3: Structure of LibriSpeech dataset

Subset	Hours	Per-Speaker Minutes	Female Speakers	Male Speakers	Total Speakers
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

Table 4: Data on Russian language corpora

Corpus	Application	Number of Speakers	Gender Distribution	Age Distribution	Specific Features	Segmentation
Mozilla Common Voice Dataset	Both	2500+	Mentioned	Mentioned	None	Word/phrase level
CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages	Synthesis	1-100	Mentioned	Not mentioned	None	Word/phrase level
Building Test Speech Dataset on Russian Language for Spoken Document Retrieval Task	Recognition	Not mentioned	Not mentioned	Not mentioned	None	None
RuSpeech	Recognition	100-500	Mentioned	Mentioned	Accent	Phonetic level
Dream Stories	Recognition	100-500	Mentioned	Mentioned	Disease	Word/phrase level
VitalVoice	Synthesis	Not mentioned	Not mentioned	Not mentioned	None	None
The National Corpus of the Russian Language	Both	Not mentioned	Mentioned	Mentioned	Multiple	Phonetic level
1,002 Hours - Russian Speech Data by Mobile Phone	Recognition	1000-2500	Mentioned	Mentioned	Command	Word/phrase level
Golos	Recognition	Not mentioned	Not mentioned	Not mentioned	None	Word/phrase level
Must-C	Recognition	2500+	Mentioned	Not mentioned	Translation	Word/phrase level
CoVoST	Recognition	100-500	Not mentioned	Not mentioned	Translation	Word/phrase level
MaSS	Recognition	Not mentioned	Not mentioned	Not mentioned	Translation	Word/phrase level
LibriSpeech	Recognition	1000-2500	Mentioned	Not mentioned	None	Word/phrase level
VoxDIY	Recognition	100-500	Not mentioned	Not mentioned	None	Word/phrase level

English Speech Corpora

Daily Talk: Spoken Dialogue Dataset for Conversational Text-To-Speech was developed for use in speech synthesis systems (Lee *et al*., 2022). All recordings are in English. Two speakers were involved in the creation, one male and one female. The amount of audio material equals 78026 seconds or 21 hours. The statistics of the corpus filling is presented in Table 5.

Back *et al*. (2020) used The Speech Commands dataset from Tensor Flow (Warden, 2018). Its main purpose is to provide a way to create and then test small models. These models determine when one word from a set of ten target words is spoken, with false positives due to background noise or unrelated speech. The language of the version of the corpus used in the paper is Malayalam (a language of southwest India), but there is also an English version. This dataset is primarily intended for training neural networks,

so the amount of material in it is not measured in hours, but in number of samples - verification samples 4890, training samples 85511, validation samples 10102.

TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo *et al*., 1993) was developed for speech recognition systems. The language of the corpus is English. 630 speakers participated in the recording. The key phrases are "10 phonetically-rich sentences". A division into 8 dialects is also present. Phoneme segmentation and transcriptions of the recordings are present.

Dataset of British English speech recordings for psychoacoustics and speech processing research: The Clarity Speech Corpus was developed (Graetzer *et al*., 2022) for purposes related to speech recognition - improving the quality of signal processing by hearing aids. Speakers of both sexes - 40 people in total - participated in the creation. The key phrases were taken from the British National Corpus (BNC) and there were 250 unique sentences for each speaker. The length of a key phrase is 7-10 words. The fields specified in the audio file description are shown in Table 6.

Table 5: DailyTalk filling statistics

Feature	Male	Female	Total
# clips	11,867	11,906	23,773
# words	124,624	126,721	251,345
total duration (s)	38902	39124	78026
mean duration/clip (s)	3.278	3.286	3.282
mean # phone/clip	29,471	29.884	29.678
# distinct words	7,220	7,329	10,160
# dialogues	2.541	2,541	2,541
mean turns/dialogue	4.670	4.686	9.356
# dialogues w/fgs	962	689	1,452

Table 6: The Clarity Speech Corpus dataset structure

Field name	Description	Example
prompt	Original prompt provided to speaker	“At the moment I never feel I’m working hard enough.”
prompt_id	BNC_ID comprising a 3 letter code followed by a 5 digit number	“G21_00436”
speaker	T followed by a 3 digit No. uniquely identifying the talker	“T037”
wavfile	T<talker No.>_<BNC_ID> wav filename	“T037_ G21_00436”
index	No.	“10”
dot	Detailed orthographic transcription	“At the moment I never feel I\\’m working hard enough”

Kim *et al*. (2022) used the LJ Speech Dataset (Ito and Johnson, 2016). The corpus was developed both for synthesis systems (for this purpose it is used in the paper under consideration) and speech recognition. The corpus contains 13100 phrases of length from 1 to 10 seconds each. The total volume of the audio material is 23 hours and 55 minutes. Texts published between 1884 and 1964 were used as sources. The language of the corpus is English.

The ReVerb Challenge (Kinoshita *et al*., 2016) used a speech corpus, the key feature of which is that all recordings were made in rooms with echoes. The contest itself was precisely to improve the quality of speech recognition on recordings with increased reverberation, i.e. echo. The language of the dataset is English. The dataset includes a training set, a development test set, and an evaluation (final) test set (discontinued on December 12, 2013). The development test kit and the final assessment data set consist of the following parts:

1. Simulated Data (Sim Data): Audio signals from the WSJCAM0 enclosure converted into Room Impulse Response (RIR), which are measured in various rooms. The recorded background noise is added to the reverberation test data with a fixed Signal-to-Noise Ratio (SNR)
2. Real data: Recordings from the MC-WSJ-AV enclosure, which consists of audio recordings recorded in a noisy and echoing room

The audio material was categorized into 1, 2 and 8 channel recordings. The exact number and gender of the speakers is not known.

Two English datasets were mentioned in this contest, namely WSJCAM0 and MC-WSJ-AV. About them further on.

WSJCAM0 Cambridge Read News (Robinson *et al*., 1995; Fransen *et al*., 1997) was developed for speech recognition systems. The structure of the corpus involves 4 samples: Training, test, and correction for both of these samples. The recording was done in mono channel mode with a sampling rate of 16 kHz. For training samples, the recording microphone was fixed on the speaker's head. For test samples - on the table in front of him. The corpus utilizes phoneme segmentation. 140 speakers participated in the recording, whose gender separation is not specified. Figure 3 shows an excerpt from the transcript attached to the audio recording.

MC-WSJ-AV (Lincoln and Zwyssig, 2012) was developed for research on speaker localization, (blind) separation and speech recognition. The textual material used was sentences that had previously also been used in the WSJCAM0 dataset. In total, about 45 speakers, male and female, were recorded in three different scenarios, namely.

```
    0     256   sil
  256    3328   f
 3328    5120   iy
 5120    6144   m
 6144    7680   ey
 7680    9216   l
 9216   10752   p
10752   11520   r
11520   11776   ax
11776   12288   d
12288   13568   y
```

Fig. 3: Transcript fragment of one of the WSJCAM0 corpus records

1. A single stationary speaker
2. Two stationary speakers interrupting each other
3. A single moving speaker

The recording was conducted on an eight-channel headset microphone.

Hernandez *et al*. (2018) developed TED-LIUM Release 3 for speech recognition systems. The corpus uses recordings of TED Talks conferences in English as the materials of the corpus. The corpus is distributed in two variations: For training models with the classical train-dev-test structure (the same as in previous versions of the corpus) and for "speaker adaptation" systems. It contains 452 hours of audio material in the form of 2351 described recordings with segmentation and transcription. The recordings are in mono-channel mode with a sampling rate of 16 kHz and a bit rate of 256 Kbit. The number and gender distribution of speakers is not specified.

VOiCES (Richey *et al*., 2018): Voices Obscured in Complex Environmental Settings was developed for speech recognition, speaker identification, speech/non-speech event detection, source localization, noise reduction, and other systems. The speech corpus contains the following data: The original audio, the translated audio, the spelling transcription of the audio, and data about the speaker. The purpose of this corpus is to promote acoustic research by providing access to acoustic data of varying complexity. The corpus contains 15 hours of audio material in the form of 3903 audio files. Speakers of both genders are said to have participated in the recording, but their number is not specified. The peculiarity is the presence of various interferences: Noise, movement of the sound source, echo. This is what is meant by "complex acoustic data". The language of the corpus is English.

VoxPopuli (Wang *et al*., 2021 and 2022) was developed for speech recognition and translation systems. The recordings of European Parliament meetings from 2009-2020 were used as a source. The corpus contains recordings in 23 languages, including English, namely:

1. 400000 hours of undescribed recordings in 23 languages
2. 1800 hours of described recordings in 16 languages
3. 17300 hours of 15x15 language-translated records in all directions
4. 29 hours of transcribed recordings in English from non-native speakers with accents (15 different accents). Table 7 shows the statistics for this section for each of the accents

The records are phonemically segmented and described. The corpus is prepared for model training and connectivity as a Python library. It is claimed that 4.3 thousand speakers participated in the recordings. Detailed statistics on the ratio of described recordings as well as the gender ratio of speakers are presented in Table 8.

Table 7: Statistics for the accentuated part of the VoxPopuli corpus

Accent	Code	Transcribed (h)	Transcribed Speakers
Dutch	en_nl	3.52	45
German	en_de	3.52	84
Czech	en_cs	3.30	26
Polish	en_pl	3.23	33
French	en_fr	2.56	27
Hungarian	en_hu	2.33	23
Finnish	en_fi	2.18	20
Romanian	en_ro	1.85	27
Slovak	en_sk	1.46	17
Spanish	en_es	1.42	18
Italian	en_it	1.11	15
Estonian	en_et	1.08	6
Lithuanian	en_lt	0.65	7
Croatian	en_hr	0.42	9
Slovene	en_sl	0.25	7

Table 8: Detailed statistics of the VoxPopuli corpus

	Unlab.	Transcribed			LM
	Hours	Hours	Speakers (F%)	Tokens	Tokens
En	24.1K	543	1313 (29.6)	4.8M	60.1M
De	23.2K	282	531 (30.6)	2.3M	50.0M
Fr	22.8K	211	534 (38.6)	2.1M	58.6M
Es	21.4K	166	305 (40.6)	1.6M	57.4M
Pl	21.2K	111	282 (23.7)	802K	13.6M
It	21.9K	91	306 (33.8)	757K	52.1M
Ro	17.9K	89	164 (27.6)	739K	10.3M
Hu	17.7K	63	143 (30.3)	431K	13.0M
Cs	18.7K	62	138 (24.9)	461K	13.5M
Nl	19K	53	221 (39.3)	488K	54.6M
Fi	14.2K	27	84 (56.8)	160K	34.5M
Hr	8.1K	43	83 (33.1)	337K	285K
Sk	12.1K	35	96 (33.8)	270K	13.3M
Sl	11.3K	10	45 (43.9)	76K	12.6M
Et	10.6K	3	29 (43.7)	18K	11.3M
Lt	14.4K	2	21 (14.8)	10K	11.5M
Pt	17.5K	-	-	-	-
Bg	17.6K	-	-	-	-
El	17.7K	-	-	-	-
Lv	13.1K	-	-	-	-
Mt	9.1K	-	-	-	-
Sv	16.3K	-	-	-	-
Da	13.6K	-	-	-	-
All	384K	1791	4295	15M	467M

GigaSpeech (Chen *et al*., 2021) was developed for speech recognition systems. The corpus comes in several variations in terms of the amount of audio material, as shown in Table 9. The largest one counts 10,000 hours. The language of the dataset is English.

The materials used were recordings from audiobooks, podcasts and YouTube videos. Hence the absence of any information on the number and gender distribution of the speakers. The structure of the dataset follows the train-dev-test scheme. Table 10 shows the number of hours for the dev and test samples, common to all variations of the corpus.

Lip Reading Sentences 3 (LRS3) Dataset (Afouras *et al*., 2018) was developed for speech and facial expression recognition system. The case materials used are recordings of TED Talks conferences in English. The number and gender distribution of speakers is not specified. The structure of the dataset follows the pretrain-train-test scheme. A key feature is that the recording is a video with a close-up of the speaker's face. The statistical distribution over the samples is presented in Table 11.

Taskmaster (Byrne *et al*., 2019) was developed for speech recognition systems. Replicas from six scenarios were used as a data source: Ordering a pizza, making an appointment at a car repair shop, ordering a cab, ordering movie theater tickets, ordering coffee, and booking a room at a restaurant. The corpus consists of text and voice parts with a separation of 7708 and 5507 dialogs, respectively. All recordings are in English. The recordings were crowdsourced, so the exact number and gender distribution of the speakers is not specified. The structure of the dataset is as follows: The dialog is divided into individual utterances and has attributes conversation Id and instruction Id. The latter describes what commands were given by the speaker to the interlocutor. Each utterance has index, speaker, text and segments fields. Segmentation in the corpus is present at the word and phrase level.

The segment fields are: Start index, end Index, text, annotation.

Table 9: Giga Speech corpus variations

Subset	Audiobook	Podcast	YouTube	Total
XL	2,655h	3,499h	8,846h	10,000h
L	650h	875h	975h	2,500h
M	260h	350h	390h	1,000h
S	65h	87.5h	97.5h	250h
XS	2.6h	3.5h	3.9h	10h

Table 10: Number of hours in dev and test samples in the Giga Speech corpus

Set	Podcast	YouTube	Total
DEV	6.3h	6.2h	12.5h
TEST	16.1h	24.2h	40.3h

Table 11: Statistical distribution over the LRS3 corpus samples

Set	# videos	# utterances	# word instances	Vocab
Pre-train	5,090	118,516	3.9M	51k
Trainval	4,004	31,982	358k	17k
Test	412	1,321	10k	2k

The latter is responsible for the presence of annotations to this segment and has a name attribute. The structure also includes the annotation ontology schema, which defines the speaker commands and the state (stage) of the dialog.

Havard *et al*. (2017) developed SPEECH-COCO for research in language acquisition, unsupervised term detection, keyword detection or semantic embedding using speech and vision as a complement to the MSCOCOCO dataset. The latter is used for image and text recognition. The corpus is structurally divided into train and validation samples and contains 616767 voiced signatures from the MSCOCOCO validation and train subsets (respectively 414113 for train and 202654 for validation). Eight speakers participated in the recording, 3 male and 5 female. 4 of them have British accents and the other 4 have American accents. The text is specifically pronounced with slight inherent hesitations, sighs and other sounds unrelated to the text. Each entry is described by its corresponding json file with metadata.

Speaking Faces (Abdrakhmanova *et al*., 2021) was developed for speech and facial expression recognition system. The language of the corpus is English. The number of speakers is 142. If there are speakers of both genders, their numbers are not specified. Structurally, the dataset is not divided into samples and is actually divided into two parts - by recording sessions. The key feature is that the recording is a combination of regular camera video, thermal camera video, and audio recording. For the most part, the purpose of this dataset has a much more significant bias toward facial expressions.

Radio Talk (Beeferman *et al*., 2019) was developed for speech recognition systems. It is intended for training models, in particular, in the description there is a manual for use with Kaldi, so the structure of the dataset implies a split into three training-dev-test samples. As a source of data are taken recordings of American radio stations. The language of the dataset is English. Each record is divided into segments, which are accompanied by a description in json format, containing the following fields:

1. Content: Transcribed speech from the fragment
2. Call sign: Call signs of the stations on which the fragment was broadcast
3. City: The city where the station is located, as specified in FCCC documents
4. State: The state where the station is located, as indicated in FCCC documents
5. Show name: The name of the broadcast, possibly this fragment
6. Signature: The initial 8 bytes of the MD5 content hash after conversion to lower case and highlighting English stop words (specifically the NLTK stop word list) to aid in duplication
7. Studio or phone: A flag indicating the source of the main sound - the phone or studio audio equipment

8. Guessed gender: The assumed gender of the speaker
9. Segment start time: Timestamp of the start of the segment
10. Segment end time: Timestamp of the end of the segment
11. Speaker id: Speaker identifier
12. Audio chunk id: Identifier of the audio fragment from which this fragment is taken (each fragment can be divided into several fragments)

EmoSpeech-Dataset (Banga *et al*., 2019) was developed for speech recognition systems. The language of the dataset is English with the presence of various Indian accents and also Indian itself. Currently the corpus operates with the keywords help, bachao, stop, no, go, yes and is under development. Unidentified commands and blank entry are also defined as key phrases. Recordings are categorized by emotion into Happy, Angry, Fearful, Calm, None. There is also a classification by ambience: Safe, sounds of gunshots, sounds of breaking glass. The audio recording has a title of the form keyword-environment-emotion-timestamp-hash.wav. 250 speakers participated in the crowdsourced data collection, collectively recording more than eight thousand individual recordings. The distribution of the number of recordings by emotion is presented in Table 12 and by command in Table 13.

The dataset assumes the use of models for training. In an experiment conducted during development, the separation into training, validation, and test samples was performed in a ratio of 70:10:20.

Organized Data on English Language Corpora

All the considered corpus for the English language were analyzed in terms of the selected parameters. Table 14 shows the collected data, which were processed.

Table 12: EmoSpeech-Dataset recording distribution by speaker

Word	Number of Samples
Calm	3826
Fearful	1630
Happy	988
Angry	1136

Table 13: EmoSpeech-Dataset recording distribution by command

Word	Number of Utterances
Help	1070
Bachao	948
Yes	877
No	787
Stop	788
Go	965
Unknown	1192

Table 14: Data on English language corpora

Corpus	Application	Number of Speakers	Gender Distribution	Age Distribution.	Specific Features	Segmentation
Mozilla Common Voice Dataset	Both	2500+	Mentioned	Mentioned	None	Word/phrase level
DailyTalk	Synthesis	1-100	Mentioned	Not mentioned	None	None
Dataset of British English speech recordings for psychoacoustics and speech processing research	Recognition	1-100	Mentioned	Not mentioned	None	Word/phrase level
TensorFlow Speech Commands	Recognition	Not mentioned	Not mentioned	Not mentioned	Command	Word/phrase level
TIMIT Acoustic-Phonetic Continuous Speech Corpus	Recognition	500-1000	Not mentioned	Not mentioned	Accent	Phonetic level
The LJ Speech Dataset	Both	Not mentioned	Not mentioned	Not mentioned	None	Phonetic level
MaSS	Recognition	Not mentioned	Not mentioned	Not mentioned	Translation	Word/phrase level
LibriSpeech	Recognition	1000-2500	Mentioned	Not mentioned	None	Word/phrase level
ReVerb Challenge	Recognition	Not mentioned	Mentioned	Not mentioned	Other	Phonetic level
WSJCAM0 Cambridge Read News	Recognition	100-500	Not mentioned	Not mentioned	None	Phonetic level
MC-WSJ-AV	Recognition	1-100	Not mentioned	Not mentioned	Other	Word/phrase level
VoxPopuli	Recognition	1-100	Mentioned	Not mentioned	Translation	Word/phrase level
TED-LIUM Release 3	Recognition	Not mentioned	Not mentioned	Not mentioned	None	Phonetic level
VOiCES:	Recognition	Not mentioned	Mentioned	Not mentioned	Multiple	Phonetic level
GigaSpeech	Recognition	Not mentioned	Not mentioned	Not mentioned	None	Word/phrase level
Lip Reading Sentences 3 Dataset	Recognition	Not mentioned	Mentioned	Not mentioned	Multimedia	Word/phrase level
Taskmaster	Recognition	Not mentioned	Not mentioned	Not mentioned	Command	Word/phrase level
SPEECH-COCO	Recognition	1-100	Mentioned	Not mentioned	Multiple	Phonetic level
SpeakingFaces	Recognition	100-500	Mentioned	Not mentioned	Multimedia	Word/phrase level
RadioTalk	Recognition	Not mentioned	Mentioned	Not mentioned	None	Word/phrase level
CrowdSpeech	Recognition	1000-2500	Not mentioned	Not mentioned	None	Word/phrase level
EMOSPEECH-DATASET	Recognition	100-500	Not mentioned	Not mentioned	Command	Word/phrase level

Some corpora have several variations. The Table presents only those of them, the records in which are in English.

Segmentation is used in most cases, but it is often not phonemic, but affects larger parts of speech. When creating a speech corpus, the gender factor is taken into account more often than the age factor. The speech characteristics of the speakers, physical or emotional state are not taken into account in these cases.

Arabic Speech Corpora

The corpus (Shareef *et al*., 2022) was developed for Arabic speech recognition systems for children with speech impairments. Thirty-eight speakers participated in the development, with a total of 770 recordings. All speakers were exclusively 7-11year old children with speech impairments (it is not specified which ones). Selected Arabic letters and numerals were used as key phrases.

The Modern Standard Arabic Phonetics for Speech Synthesis corpus was developed by Halabi (2016) for speech synthesis systems. One male speaker participated in the creation. Special attention is paid to phonetic features of the Arabic language - accentuation, prosody, hemination, nasalization, accentuation, diphthongization. Accordingly, segmentation in-to phonemes is also present. HTK version 3.4.1 was used for automatic segmentation. The total duration of the audio material is 3.7 hours.

Masc: Massive Arabic Speech Corpus (Al-Fetyani *et al*., 2023) was developed for speech recognition systems. Videos from YouTube.com were used as a data source, so there is no information about the exact number of speakers and their gender separation. The structure of the dataset tables is as follows: Channels_quality.csv (channel description: Channel_url, channel_name, channel_id, quality), audios (all audio materials), subsets (all sub tables, which are divided into sampling, noise and data type - metadata or not), subtitles (raw subtitles extracted from the video). Regular data tables from subsets have video_id, start, end, duration, text fields. Metadata tables have fields video_id, category, video_duration, channel_id, country, dialect, gender, transcript_duration. In both sub-tables video_id is a foreign key referencing the video ID from the audios table. The number of unique words in the dataset is 12 million. The cumulative duration of the audios is 1000 hours.

Mohammed (2016) was developed for speech recognition systems. Recordings of Quran recitation in Arabic are used as input data. The number and gender distribution of speakers is not specified. There is no partitioning into samples. The recordings are stored separately for each speaker and refer to a single json file with a record of all ayats and surahs, which is essentially a description for each audio file.

MediaSpeech (Kolobov *et al*., 2021) was developed for speech recognition systems. This dataset includes short fragments of speech automatically extracted from videos. The captured videos are available on YouTube and are transcribed manually with pre- and post-processing. Media Speech contains recordings with a total duration of 10 hours of oral speech in each of the following languages: French, Arabic, Turkish and Spanish. The number and gender distribution of speakers is not specified.

Tunisian_MSA (2016) was developed for speech recognition systems. The corpus is de-signed to train acoustic models for pronunciation modeling in Arabic language learning applications: Its structure, accordingly, consists of three train-dev-test samples. The corpus is relatively small: 11.2 hours of audio material. Three men from Libya and one woman from Tunisia participated in the recording.

Organized Data on Arabic Language Corpora

All the considered corpus for the Arabic language were analyzed in terms of the selected parameters. Table 15 shows the collected data, which were further, processed. Some corpora have several variations.

Table 15: Data on Arabic language corpora

Corpus	Application	Number of Speakers	Gender Distribution	Age Distribution	Specific Features	Segmentation
Towards Developing Impairments Arabic Speech Dataset Using Deep Learning	Recognition	1-100	Not mentioned	Mentioned	Speech defect	Phonetic level
Modern Standard Arabic Phonetics for Speech Synthesis	Synthesis	1-100	Mentioned	Not mentioned	None	Phonetic level
Masc: Massive Arabic Speech Corpus	Recognition	Not mentioned	Not mentioned	Mentioned	Accent	Word/phrase level
CoVoST	Recognition	100-500	Not mentioned	Not mentioned	Translation	Word/phrase level
Mohammed	Recognition	Not mentioned	Not mentioned	Not mentioned	None	Word/phrase level
MediaSpeech	Recognition	Not mentioned	Not mentioned	Not mentioned	None	None
Tunisian_MSA	Recognition	1-100	Mentioned	Not mentioned	None	None

The table presents only those of them, the records in which are in Arabic. Segmentation is used in most cases, but it is often not phonemic, but affects larger parts of speech. When creating a speech corpus, the gender factor is taken into account more often than the age factor. The speech characteristics of the speakers are not taken into account in the considered cases. Physical and emotional states are also not taken into account.

Speech Corpora in Other Languages

The corpus (Babirye *et al*., 2022) was developed for automatic speech recognition systems due to the lack of such systems for specific African languages. Languages: Luganda, Nkori-Kiga, Acholi, Masaba, Swahili (East African languages). Key phrases were taken from open sources with further adaptation. The length of the key phrase is less than 14 words. The amount of audio material equals 406 hours.

The corpus (Black, 2019) was developed for speech recognition, processing and synthesis systems. A total of 699 languages is claimed for the most part from the southern hemi-sphere. Key phrases were collected from various audio books. The amount of audio material is claimed to be an average of 20 hours per language.

Open-Source Magic Data-RAMC: A Rich Annotated Mandarin Conversational (RAMC) Speech Dataset (Yang *et al*., 2022) was developed for speech recognition systems. All recordings are in Mandarin Chinese. Recordings were made in a 20 m2 room with a reverberation time of less than 0.4 s. There were 663 speakers involved in the production, 295 of which were female and 368 males. There is also a division into South Chinese (329) and North Chinese speakers (334) due to the difference in pronunciation in the two regions. The corpus is based on randomized speech. The amount of audio material equals 180 hours.

The corpus (Tran and Ibrahim, 2020) is the FPT Open Speech Data (FOSD) (Tran, 2020). This dataset was developed for speech recognition systems. It is stated that it can be used to recognize: Speaker's gender, mood, intention, speech signal onset, as well as for speech synthesis and record noise removal. The language of the recording is Vietnamese. There are a total of 25921 recordings, giving a total of about 30 hours of audio material.

The corpus (Takeuchi *et al*., 2017) was developed for a speech and gesture recognition system. All recordings are in Japanese. Two male speakers of 25 years of age participated in the creation. A total of 1049 sentences were recorded, totaling 298 minutes of audio material. The paper puts more emphasis on poses and gestures.

WenetSpeech (Zhang *et al*., 2022) was developed for speech recognition systems. Podcast recordings and YouTube videos in Mandarin Chinese were used as input data. More than ten thousand hours of audio data with accurate descriptions and another 2400 hours of audio with inaccurate descriptions are claimed. On the official Internet resource of the corpus, it is stated that during the development the authors relied on the experience of creating the Giga Speech dataset, which has already been mentioned earlier. The number and gender distribution of the speakers is not specified. Also, the records in the corpus are divided into 10 categories according to the principle of the topic discussed.

TUDA (Radeck-Arneth *et al*., 2015) was developed for speech recognition systems. The language of the corpus is German. The dataset is mainly aimed at training the Kaldi model, so it has a classical structure in the form of train-dev-test samples. Recordings were made in parallel from three microphones (Microsoft Kinect, Yamaha, Samson). The amount of material for each microphone amounted to 36 hours, which were distributed over the samples in the ratio of 31-2.5-2.5. All recordings are accompanied by a file with a description of the spoken text. The recordings involved 180 speakers, 150 of whom were men and 30 women.

The corpus (Kjartansson *et al*., 2020) was developed for speech recognition systems. It consists of three separate datasets for Galician, Catalan and Basque. 132 speakers participated in the development, recording a total of 33 hours of audio material. A detailed distribution of the content is shown in Table 16. The key phrases used are the so-called "typical sentences", in which a large number of proper names were used, which are of the form "name went from $place_a$ to $place_b$ to *time*" or similar, where the italicized words were replaced by proper names.

Table 16: Open-Source article High Quality Speech Datasets for Basque, Catalan and Galician dataset filling statistics.

			Tokens					Chars				Audio Duration	
Language	Gender	Lines	min	max	avg	Total	Unique	min	max	avg	Speakers	Total [h:m:s]	Average
Basque	F	3,858	1	20	8.0	30,901	8,583	17	156	58.1	29	7:26:36	6.77
	M	3,278	1	18	8.0	26,383	8,030	23	129	58.3	23	6:36:00	7.25
Catalan	F	2,321	2	24	10.5	24,385	6,586	17	142	59.5	20	5:24:00	8.38
	M	1,919	2	29	10.6	20,261	6,514	28	141	60.8	16	4:01:12	7.53
Galician	F	4,264	3	28	11.6	49,674	6,530	18	174	68.3	34	7:40:12	6.48
	M	1,324	4	28	11.7	15,462	4,336	20	186	69.4	10	2:38:24	7.19
Total	-	16,963	-	-	-	167,066	-	-	-	-	132	33:35:19	-

The Norwegian Parliamentary Speech Corpus (Solberg and Ortiz, 2022) was developed for speech recognition systems. All recordings of the corpus are in Norwegian. 41 recordings of parliamentary debates were used as sources. The processing resulted in 140 hours of audio material containing a total of 1.2 million words and 64531 sentences. The number of speakers is not specified, but the gender ratio is 38.3% female and 61.7% male.

ROBIN Technical Acquisition Speech Corpus (Paiş *et al*., 2021) was developed for speech recognition systems. The language of the corpus is Romanian. Despite the use of model learning, the dataset is not divided into samples and represents a single structure. The recordings are represented by three files: An audio in WAV format, an annotation in CoNLL-U format and a text paraphrase of the spoken phrase in TXT format. The annotation includes segmentation at the word level (with part-of-speech description) and phrase level, but not at the phoneme level. Statistics on the used textual material is presented in Table 17. Statistical data on audio materials are presented in Table 18.

Table 19 also provides statistics on the number of parts of speech found in the entire dataset.

Table 17: ROBINTASC text usage statistics

Static	Value
Number of text files	711
Total text size	57 Kb
Maximum text size	122b
Minimum text size	3b
Average text size	81.8b
Number of tokens	11, 927
Unique tokens	222
Unique lemmas	191
Hapax legomena	58

Table 18: ROBINTASC audio usage statistics

Static	Value
Number of WAV files	3786
Total duration	6h25m03s
Maximum duration	1.02s
Minimum duration	12.91s
Average duration	6.10s
Total size	1.89Gb
Sample rate	44.1KHz
Channels	1
Encoding	Signed Int16 PCM

Table 19: ROBINTASC audio usage statistics

Tag	Occurrence	# Unq. Lemmas
NOUN	2,675	66
ADJ	1,698	32
DET	1,211	7
NUM	1,089	21
ADP	919	9
VERB	558	29
ADV	514	14
PRON	485	5
AUX	467	3

Six speakers took part in the recording: Three men and three women. Distribution by the number of recordings for each speaker, as well as their age affiliation is presented in Table 20.

ClovaCall (Ha *et al*., 2020) was developed for speech recognition systems. The language of the corpus is Korean. The source materials used were dialog recordings of a total of 11,000 people of different genders and an AI autoresponder. The recordings are accompanied by a textual transcript and consist of a small sentence separated from the dialog.

In total, the corpus contains approximately 112 thousand records. The corpus is intended for model training and is structurally divided into three raw-train-test samples. The distribution by the number of hours and records in each sample is presented in Table 21. In this case, hours are subdivided into processed and untreated.

Organized Data on Other Language Corpora

The above-described speech corpora were analyzed with respect to the parameters under study. Table 22 shows the collected data which were further processed.

Table 20: ROBINTASC speaker statistics

Spk	Gender	Age	Audio Files
1	M	40-50	233
2	M	30-40	711
3	M	20-30	711
4	F	30-40	711
5	F	40-50	709
6	F	40-50	711

Table 21: ClovaCall distribution by sample

Dataset	Number	Hour (raw/ clean)
Raw	81,222	125 / 67
Train	59,662	80 / 50
Test	1,084	1.660 / 0.88

Results

The data shown in Tables 4, 14, 15 and 22 have been processed. This section will provide statistics for all the parameters considered. When making calculations, variations of corpora in different languages are counted as several different corpora. Also, if a corpus has several variations in languages other than Russian, English, or Arabic, as well as in one or more of them, then only the variations in the above languages are counted. That is, if there were variations, for example, in German, French and English, only the English version will be included in the overall statistics. A total of 53 datasets were considered, taking into account the rules described above. The statistics by language are shown in Fig. 4.

It was found that English is represented in one way or another in most of the datasets studied. This distribution

was expected because English is the language of international communication, and because of the generally higher level of development of English-language science. Also in the aggregate, the corpus for Russian and Arabic together turned out to be as numerous as for English. At the same time, significantly fewer datasets were found for Arabic: Half as many as for Russian, and three times as few as for English.

The statistics by application are shown in Fig. 5. These statistics are generally not intended to draw any conclusions and have a representative function, as the aim of the paper was to investigate primarily speech recognition-related corpora.

The statistics by number of speakers are shown in Fig. 6. It was found that in almost half of the cases this parameter is not specified. By and large, this is due to the fact that speech corpora are most often created using crowdsourcing, in which it is difficult to track the number of speakers. By the same principle, more than two thirds of the categories 1-100 and 100-500 speakers are taken together - if crowdsourcing was not used in the work, then acquaintances and colleagues were involved in the number of speakers that can be processed by the research team involved in the creation of the dataset.

The statistics by segmentation level are shown in Fig. 7.

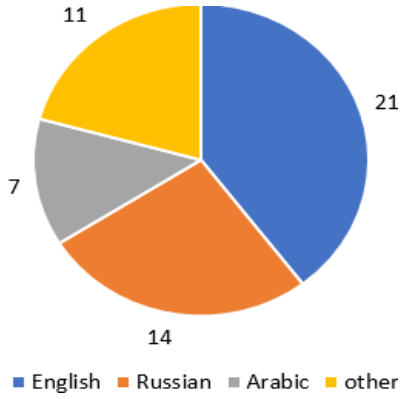

Fig. 4: Statistics of reviewed datasets by language

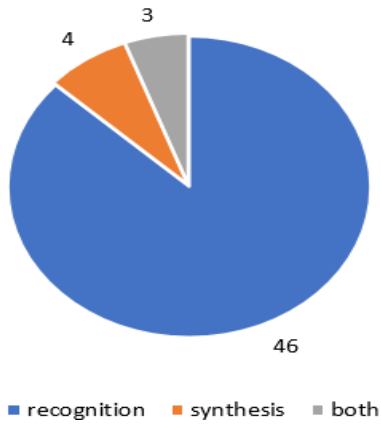

Fig. 5: Statistics of reviewed datasets by application

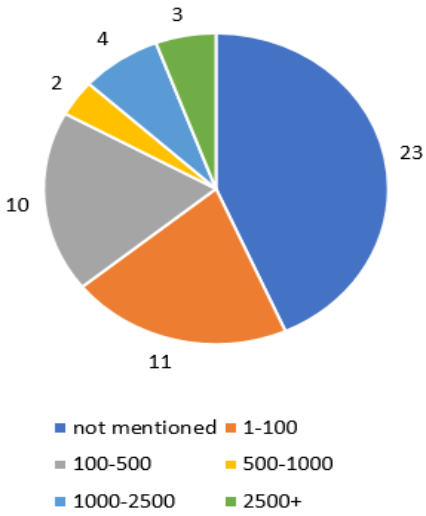

Fig. 6: Statistics of reviewed datasets by number of speakers

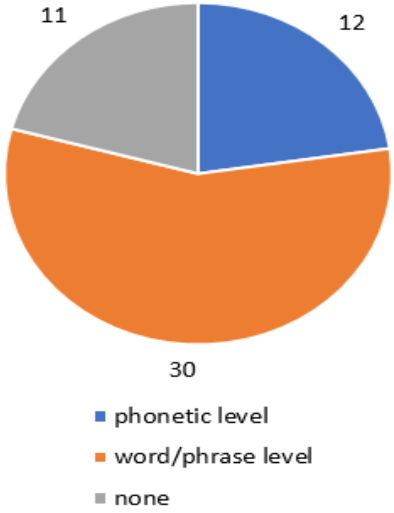

Fig. 7: Statistics of reviewed datasets by segmentation level

It was found that the most common level of segmentation was into large speech units. This parameter depended directly on the characteristics of the dataset. Phoneme segmentation was used where the goal implied greater accuracy, such as in datasets with high noise or speakers with accents. The absence of segmentation is mainly in small cases.

The statistics by age and gender distribution are shown in Fig. 8.

It was found that gender was present in more than half of the corpora, while the age of the speakers was reported in less than 20% of the corpora examined. Following from this, it can be assumed that age is of much less interest to the creators of speech corpora as a factor influencing vocal characteristics. Also, this parameter is again influenced by crowdsourcing, where it is often problematic to specify the age and/or gender of the speaker.

The statistics by specific features presence are shown in Fig. 9.

About half (26 out of 53) of the datasets contained additional specific fields. Given the fact that the selection of datasets did not imply any quota for this type of corpus, it can be argued that there is a demand for task-specific solutions. Therefore, the presence of some additional specific division may increase the value of the speech corpus produced.

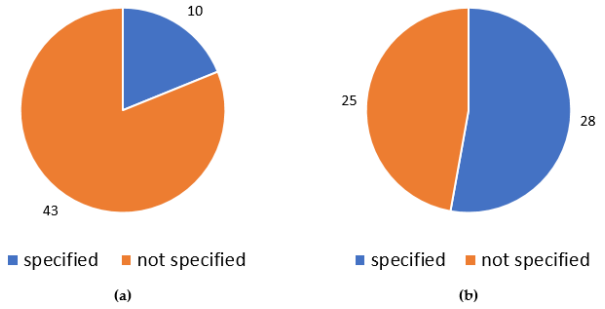

Fig. 8: Statistics of reviewed datasets by age (a) and gender (b) distribution presence

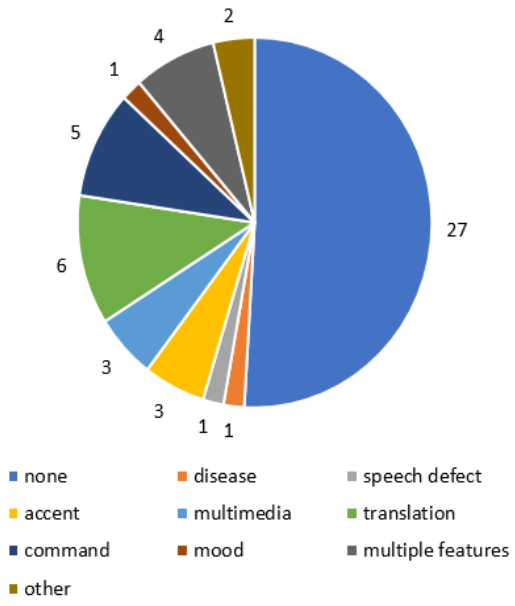

Fig. 9: Statistics of reviewed datasets by specific features presence

Discussion

Fifty different speech blocks were studied, or 53 if language variations are additionally taken into account. The structure of each block has been reviewed to give an idea of how data is stored in it.

As a result, the following parameters of the speech signal can be distinguished:

1. The gender of the speaker. Male and female voices are different and sound different - they have different timbre, pitch, and other vocal parameters. In this regard, taking this factor into account is not uncommon when creating a speech corpus
2. The age of the speaker. As with the previous parameter, age is an important factor in the sound of the voice, affecting many speech parameters. As the speaker's age increases, the parameters of his speech change significantly
3. The presence of specific features. There are data sets in which speech is presented in a pure, "ideal" form. However, they are not enough if it is necessary to conduct research, for example, in the field of medicine, affecting the speech of people with articulation disorders. For these purposes, special datasets are being developed that take into account some additional factors. However, they are not enough if it is necessary to conduct research, for example, in the field of medicine, affecting the speech of people with articulation disorders. Special data sets are being developed for these purposes, taking into account some additional factors

Parameters for describing speech corpora:

1. Languages. The linguistic component of the corpus largely depends on this parameter - segmentation by phonemes and, in rare cases, by parts of speech, as well as transcription
2. Areas of application. The main ones are speech recognition and synthesis. Depending on the specific corpus, the purpose of its use is determined: The corpus for speech synthesis systems is poorly suited for recognition and vice versa. There are also multi-purpose enclosures that are suitable for both tasks. Speech synthesis corpora also seem to show differences in structure and content depending on the purpose of the corpus
3. The number of speakers. This parameter determines the variety of voices represented in the corpus, and therefore its versatility. Some authors, however, do not specify such data, which does not complicate the work with the case as a whole, but makes it difficult to understand the amount of work done and the parameters of the output product
4. Segmentation. Its use in datasets allows you to study the sound of a single phonetic element, word or phrase. In the case of phoneme segmentation, the recording is often accompanied by transcription

Conclusion and Future Work

The purpose of this article was to study the structure of various speech corpora for different languages. The following parameters were evaluated: Language, field of application, number, gender and age division of speakers, segmentation level, specific directions.

The corpora for Russian and Arabic are represented in smaller number than for English. The corpora developed for speech recognition systems exist in sufficient numbers to be investigated, and their structure is often not too different from corpora for speech synthesis systems, as evidenced, for example, by the presence of multifunctional corpora. Speaker selection is predominantly done in two scenarios: Crowdsourcing and "in-house selection". This in turn affects the number of speakers. Segmentation is used in most cases, but it is often not phonemic but affects larger parts of speech. Gender is more often considered than age when creating a speech corpus. Specific directions are in demand and theoretically their presence can increase the value of a speech corpus. In future work, when creating the corpora, it will be necessary to take into account the highlighted features of the speakers.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

This work was supported by the Ministry of Science and Higher Education of Russia, Government Order for 2023-2025, project no. FEWM-2023-0015 (TUSUR).

Author's Contributions

Vladimir Igorevich Fedoseev: Conceptualization, validation, resources, data curation, writing original draft preparation.

Anton Aleksandrovich Konev: Conceptualization, validation, data curation, writing review and edited.

Natalia Sergeevna Repyuk: Conceptualization, data curation, writing review and edited.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript, and no ethical issues involved.

References

Abdrakhmanova, M., Kuzdeuov, A., Jarju, S., Khassanov, Y., Lewis, M., & Varol, H. A. (2021). SpeakingFaces: A Large-Scale Multimodal Dataset of Voice Commands with Visual and Thermal Video Streams. *Sensors*, *21*(10), 3465. https://doi.org/10.3390/s21103465

Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition. *Computer Vision and Pattern Recognition*, *1*, 1–8. https://doi.org/10.48550/arXiv.1809.00496

Al-Fetyani, M., Al-Barham, M., Abandah, G., Alsharkawi, A., & Dawas, M. (2023). MASC: Massive Arabic Speech Corpus. *2022 IEEE Spoken Language Technology Workshop (SLT)*, 1–6. https://doi.org/10.1109/slt54892.2023.10022652

Arlazarov, V. L., Bogdanov, D. S., Krivnova, O. F., & Ya, A. (2004). Creation of Russian Speech Databases: Design, Processing, Development Tools. *ISCA Archive*, 650–656.

Babirye, C., Nakatumba-Nabende, J., Franics, J., Mukiibi, J., Katumba, A., Ogwang, R., Sentanda, M., Wanzare, L., & David, D. (2022). *Building Text and Speech Datasets for Low Resourced Languages: A Case of Languages in East Africa*. 1–6.

Back, M.-K., Yoon, S.-W., & Lee, K.-C. (2020). GAN-based Augmentation for Populating Speech Dataset with High Fidelity Synthesized Audio. *IEEE Xplore Digital Library*, 1267–1269. https://doi.org/10.1109/ictc49870.2020.9289283

Banga, S., Upadhyay, U., Agarwal, Piyush, Sharma, S., & Mukherjee, M. (2019). Indian EmoSpeech Command Dataset: A dataset for emotion-based speech recognition in the wild. *Audio and Speech Processing*, *1*, 1–7. https://doi.org/10.48550/arXiv.1910.13801

Beeferman, D., Brannon, W., & Roy, D. (2019). RadioTalk: A Large-Scale Corpus of Talk Radio Transcripts. *Interspeech 2019*, 564–568. https://doi.org/10.21437/interspeech.2019-2714

Black, A. W. (2019). CMU Wilderness Multilingual Speech Dataset. *Speech and Signal Processing (ICASSP)*, 5971–5975. https://doi.org/10.1109/icassp.2019.8683536

Boito, M. Z., William, H., Garnerin, M., Eric Le, F., & Laurent, B. (2019). MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6486–6493. https://doi.org/https://doi.org/10.48550/arXiv.1907.12895

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., Yavuz, S., Dubey, A., Kim, K.-Y., & Cedilnik, A. (2019). Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. *Association for Computational Linguistics (ACL Anthology) – Accessible*, 4516–4525. https://doi.org/10.18653/v1/d19-1459

Cattoni, R., Di Gangi, M. A., Bentivogli, L., Negri, M., & Turchi, M. (2021). MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, *66*, 101155. https://doi.org/10.1016/j.csl.2020.101155

Chen, G., Shiyin, C., Gong, W., Jinyu, D., Wei-Qiang, Z., Changliang, W., Dong, Y., Daniel, P., Jan, T., & Jie, Zhang. (2021). GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of. *Computer Science Sound*, *1*, 3670–3674. https://doi.org/https://arxiv.org/abs/2106.06909

Datatang. (2022). Hours - Russian Speech Data by Mobile Phone_Data Products. *Datatang Official Dataset Repository*. https://www.datatang.ai/datasets/976

Fransen, J., Pye, D., Robinson, T., Woodland, P., & Young, S. (1997). WSJCAM0 Corpus and Recording Description. *Academia*, 831–834.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). *DARPA TIMIT:* https://doi.org/10.6028/nist.ir.4930

Graetzer, S., Akeroyd, M. A., Barker, J., Cox, T. J., Culling, J. F., Naylor, G., Porter, E., & Viveros-Muñoz, R. (2022). Dataset of British English speech recordings for psychoacoustics and speech processing research: The clarity speech corpus. *Data in Brief*, *41*, 107951. https://doi.org/10.1016/j.dib.2022.107951

Ha, J.-W., Nam, K., Kang, J., Lee, S.-W., Yang, S., Jung, H., Kim, H., Kim, E., Kim, S., Kim, H. A., Doh, K., Lee, C. K., Sung, N., & Kim, S. (2020). ClovaCall: Korean Goal-Oriented Dialog Speech Corpus for Automatic Speech Recognition of Contact Centers. *Interspeech 2020*. Interspeech 2020. https://doi.org/10.21437/interspeech.2020-1136

Halabi, N. (2016). *Modern standard Arabic phonetics for speech synthesis*.

Havard, W., Besacier, L., & Rosec, O. (2017). SPEECH-COCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO Data Set. *ISCA Archive (International Speech Communication Association)*, 42–46. https://doi.org/10.21437/glu.2017-9

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., & Estève, Y. (2018). *TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation*. 198–208. https://doi.org/10.1007/978-3-319-99579-3_21

Ito, K., & Johnson, L. (2016). The LJ Speech Dataset. *Official Dataset Page*. https://keithito.com/LJ-Speech-Dataset/

Karpov, N., Denisenko, A., & Minkin, F. (2021). Golos: Russian Dataset for Speech Research. *Interspeech 2021*, 1419–1423. https://doi.org/10.21437/interspeech.2021-462

Kibrik, A. A., Podlesskaya, V. I., Nikolay A., N. A., Litvinenko, A. O., Buryakov, M. L., Ilyina, E. I., Yefimova, Z. V., Hurshudyan, V. G., Vardinyan, E. G., & Orlova, K. V. (2009). Spoken Corpora of Russia. *Published in Lecture Notes in Computer Science*.

Kim, M., Jeong, M., Choi, B. J., Ahn, S., Lee, J. Y., & Kim, N. S. (2022). Transfer Learning Framework for Low-Resource Text-to-Speech using a Large-Scale Unlabeled Speech Corpus. *Interspeech 2022*, 788–792. https://doi.org/10.21437/interspeech.2022-225

Kinoshita, K., Delcroix, M., Gannot, S., P. Habets, E. A., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A., & Yoshioka, T. (2016). A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 1–28. https://doi.org/10.1186/s13634-016-0306-6

Kjartansson O., Gutkin A., Butryna A., Demirsahin I., Rivera C. (2020). Open-Source High Quality Speech Datasets for Basque, Catalan and Galician. Proc. of 1st Joint Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL) Workshop (SLTU-CCURL 2020), European Language Resources Association (ELRA), Marseille, France, 21-27.

Kolobov, R., Okhapkina, O., Omelchishina, O., Platunov, A., Bedyakin, R., Moshkin, V., Menshikov, D., & Mikhaylovskiy, N. (2021). MediaSpeech: Multilanguage ASR Benchmark and Dataset. *Audio and Speech Processing*, *1*, 1–12. https://doi.org/10.48550/arXiv.2103.16193

Lee, K., Park, K., & Kim, D. (2023). *DailyTalk: Spoken Dialogue Dataset for Conversational Text-to-Speech*. 1–5. https://doi.org/10.1109/icassp49357.2023.10095751

Lincoln, M., & Zwyssig, E. (2012). MC-WSJ-AV. *Centre for Speech Technology Research*.

Mohammed, S. (2016). Openslr.Org Dataset. Mohammed Speech Dataset. *Openslr.Org*. https://www.openslr.org/46/

Paiş, V., Ion, R., Avram, A.-M., Irimia, E., Mititelu, V. B., & Mitrofan, M. (2021). Human-Machine Interaction Speech Corpus from the ROBIN project. *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 91–96. https://doi.org/10.1109/sped53181.2021.9587355

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *IEEE Xplore*, 5206–5210. https://doi.org/10.1109/icassp.2015.7178964

Park, K., & Mulc, T. (2019). CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages. *Interspeech 2019*, 1566–1570. https://doi.org/10.21437/interspeech.2019-1500

Pavlichenko, N., Stelmakh, I., & Ustalov, D. (2021). CrowdSpeech and Vox DIY: Benchmark Dataset for Crowdsourced Audio Transcription. *Virtual Conference (Hosted by NeurIPS, Traditionally Vancouver, Canada)*, 1–11. https://doi.org/10.48550/arXiv.2107.01091

Prodeus, A. N. (2013). Speech Corpora: Creation and Problems. *Electrical and Computer Systems. ISSN*, *9*(85), 2221–3805. https://doi.org/https://eltecs.op.edu.ua/index.php/journal/article/view/1313

Radeck-Arneth, S., Milde, B., Lange, A., Gouvêa, E., Radomski, S., Mühlhäuser, M., & Biemann, C. (2015). *Open Source German Distant Speech Recognition: Corpus and Acoustic Model*. 480–488. https://doi.org/10.1007/978-3-319-24033-6_54

Richey, C., Barrios, M. A., Armstrong, Z., Bartels, C., Franco, H., Graciarena, M., Lawson, A., Nandwana, M. K., Stauffer, A., van Hout, J., Gamble, P., Hetherly, J., Stephenson, C., & Ni, K. (2018). Voices Obscured in Complex Environmental Settings (VOiCES) Corpus. *Interspeech 2018*, 1566–1570. https://doi.org/10.21437/interspeech.2018-1454

Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S., Woodland, P., & Young, S. (1995). WSJCAM0 Cambridge Read News. *Dataset Release (Part of the WSJCAM0 Project, Based on Wall Street Journal Read Speech)*. https://doi.org/https://catalog.ldc.upenn.edu/LDC95S24

Saito, K., Uhlich, S., Fabbro, G., & Mitsufuji, Y. (2021). Training Speech Enhancement Systems with Noisy. *Arxiv*, *6*, 1–5. https://doi.org/10.48550/arXiv.2105.12315

Shareef, S. R., & Al-Irhayim, Y. F. (2022). Towards developing impairments arabic speech dataset using deep learning. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*(3), 1400. https://doi.org/10.11591/ijeecs.v25.i3.pp1400-1405

Solberg, P., & Ortiz, P. (2022). The Norwegian Parliamentary Speech Corpus. *Computation and Language*, 1–10. https://doi.org/https://arxiv.org/abs/2201.10881

The National Corpus of the Russian Language. (2004). *Official Website*.

Takeuchi, K., Kubota, S., Suzuki, K., Hasegawa, D., & Sakuta, H. (2017). *Creating a Gesture-Speech Dataset for Speech-Based Automatic Gesture Generation*. 198–202. https://doi.org/10.1007/978-3-319-58750-9_28

Tatarinova, A., & Prozorov, D. (2018). Building Test Speech Dataset on Russian Language for Spoken Document Retrieval Task. *IEEE Xplore*, 1–4. https://doi.org/10.1109/ewdts.2018.8524598

Tran, D. C., & Ibrahim, R. (2020). On the Identification of FOSD-based Non-zero Onset Speech Dataset. *IEEE Xplore*, 108–110. https://doi.org/10.1109/scored50371.2020.9251018

Tran, D.-C. (2020). FPT Open Speech Dataset (FOSD) - Vietnamese. *Mendeley Data*. https://doi.org/https://data.mendeley.com/datasets/k9sxg2twv4/4

Tunisian_MSA. (2016). OpenSLR (Open Speech and Language Resources) official repository. *Openslr.Org*. https://www.openslr.org/46/

Wang, C., Pino, J., & Gu, J. (2020). CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), Marseille, France. Published by the European Language Resources Association (ELRA)*. https://doi.org/https://aclanthology.org/2020.lrec-1.517/

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021). VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 993–1003. https://doi.org/10.18653/v1/2021.acl-long.80

Warden, P. (2018). Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *Computation and Language*, 1–6. https://doi.org/10.48550/arXiv.1804.03209

Yang, Z., Chen, Y., Luo, L., Yang, R., Ye, L., Cheng, G., ... & Yan, Y. (2022). Open source magicdata-ramc: A rich annotated mandarin conversational (ramc) speech dataset. *arXiv preprint arXiv:2203.16844*. https://doi.org/10.48550/arXiv.2203.16844

Zhang, B., Lv, H., Guo, P., Shao, Q., Yang, C., Xie, L., Xu, X., Bu, H., Chen, X., Zeng, C., Wu, D., & Peng, Z. (2022). WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6182–6186. https://doi.org/10.1109/icassp43922.2022.9746682