International Journal of Scientific Advances
ISSN: 2708-7972
Volume: 3 | Issue: 4 | Jul - Aug 2022 Available Online: www.ijscia.com
DOI: 10.51542/ijscia.v3i4.17

Extended ArmSpeech: Armenian Spoken Language Corpus

Varuzhan H. Baghdasaryan

National Polytechnic University of Armenia
Armenia, Yerevan, Romanos Melikian 6/1

*Corresponding author details: Varuzhan H. Baghdasaryan; varuzh2014@gmail.com

ABSTRACT

The first paper of ArmSpeech presented an annotative native Armenian speech corpus, its data collection, preprocessing and annotation processes, corpus structure and statistics. The main reason for ArmSpeech creation is to increase Armenian language research resources because according to research there are no free or paid Armenian speech corpora for speech-to-text, text-to-speech and language research. From an NLP perspective, the Armenian language is a low-resourced language despite the fact that The Armenian language is an independent branch of the Indo-European language family and the native language of 12-15 million people. ArmSpeech corpus can be used in natural language processing (NLP) research. The first release of the corpus mainly contains audio clips extracted from free-to-use audiobooks. The total duration of audio clips is 11.77 hours. ArmSpeech's first release corpus includes 6206 audio clips of multiple speakers of any age, gender and accent. This paper intends to present the ArmSpeech extended version, which is a continuation of the previous work, includes an annotated Armenian speech, and the recording process is based on the volunteer's voice donation principle. The paper also introduces necessary data collection, pre-processing, recording and annotation stages, final results and statistics of the corpus. The material (text) needed for the recording was collected from the articles on Armenian news websites about lifestyle, culture, sport and politics. Recording was done by 1 female and 3 male volunteers whose native language is Armenian. The total duration of the data included in the second release is approximately 4 hours and along with the first release, the ArmSpeech corpus becomes 15.7 hours.

Keywords: Armenian speech corpus; speech recognition; speech-to-text; speech synthesis; text-to-speech; spoken language identification

INTRODUCTION

Before referring to the second version of ArmSpeech, it is necessary to mention the definition and significance of speech corpora. A speech corpus is a database of speech audio files and corresponding text transcriptions [1]. In natural language processing (NLP) [2] tasks such as speech recognition [3], speech synthesis [4] or spoken language identification [5], speech corpora are used to create acoustic models [6]. Also, speech corpora can be useful for linguistics research (phonetic, conversation analysis, dialectology and other fields).

Before discussing the extended version of ArmSpeech it is preferable to briefly recap the first release of ArmSpeech [7]. The first release is a multi-speaker speech corpus consisting of 14 sections. The First 13 sections contain speeches collected from free-to-use audiobooks. The last 14th section contains a speech about numerous real-life situations. The first release contains 62.37% or over 7.3 hours of male speech and 37.63% or over 4.4 hours of female speeches (11 male and 3 female voices). Also, in ArmSpeech's first paper [7] can be found short descriptions and research of related works for other languages.

The extended version of ArmSpeech mainly includes a speech about sport, lifestyle, culture and politics. The creation process was based on volunteers' voice donations. The donation process was done in a noise-protected room using a noise-reduction microphone. For annotation used the same GUI python application which used during ArmSpeech's first version last section annotation.

All audio clips do not contain leading and trailing silent parts. Audio clips were additionally verified to avoid annotation errors. The audio clips specifications are the same as in ArmSpeech's first version:

- mono-channel 16-bit WAV (lossless compression) files,
- 16000 Hz sampling rate,
- 256 kbps bit rate Audio.

DATA COLLECTION AND PRE-PROCESSING

Annotation data for speech corpus was collected (scraped) from Armenian news websites by using software tools based on web scraping technologies [8, 9]. The main articles included in the collected data are about lifestyle, culture, sport and politics. Data scraped from websites is consist of sentences. These sentences were split into parts by the punctuation marks ":" and ",", and because of this occasionally clips contain not meaningful phrases but groups of words. During the collected data normalization [10] stage all numbers and dates were replaced by their corresponding textual representation, all symbols and punctuation marks except for Armenian letters were removed, and acronyms were not manually substituted with their expanded forms.

ANNOTATION

The speech corpus recording and annotation processes can be automated by using software and hardware toolkits. Such toolkits are annotation software, noise-reduction hardware, microphones, etc.

After data collection and pre-processing, the scraped data was simply annotated by the GUI application.

Annotation program GUI is consisting of a display area which shows normalized sentences, and 3 buttons: “Repeat”, “Invalid” and “Next”. The full description can be found in ArmSpeech’s first release paper [7].

The annotation program use voice activity detection technology to detect speech by simply distinguishing between silence and speech. This is done by using Python free “webrtcvad” module, which is a python interface to the WebRTC [11] Voice Activity Detector (VAD) [12] developed by Google. The application determines voice activity by a ratio of not null and null frames in 300 milliseconds (see figure 1). The portion of not null frames in given milliseconds must be equal to or greater than 75%.

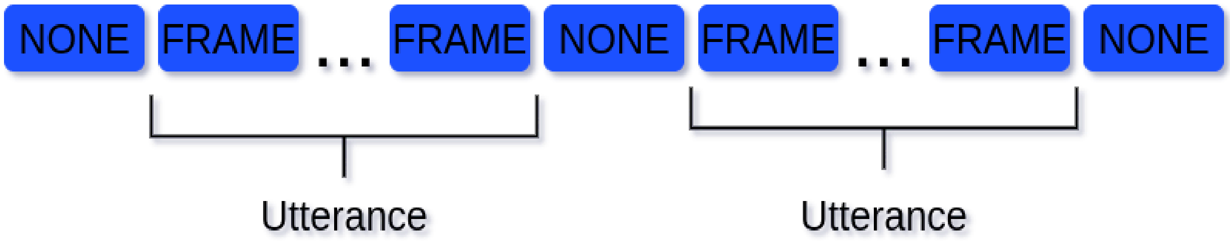

FIGURE 1: The principle of voice activity detection.

It starts to record when the volunteer starts to read the normalized text and ends recording when the speaker stops speaking. Then the application automatically stores audio frames into file in WAV format without including leading and trailing parts in the recorded audio clip.

STATISTICS OF THE CORPUS

Figure 2 shows the ratio of male and female speeches in ArmSpeech along with the extended version. At this time again male speeches are more than female speeches. Only the data for extension contains 68.3% male (2.15 times greater than female speeches) and 31.7% female speeches.

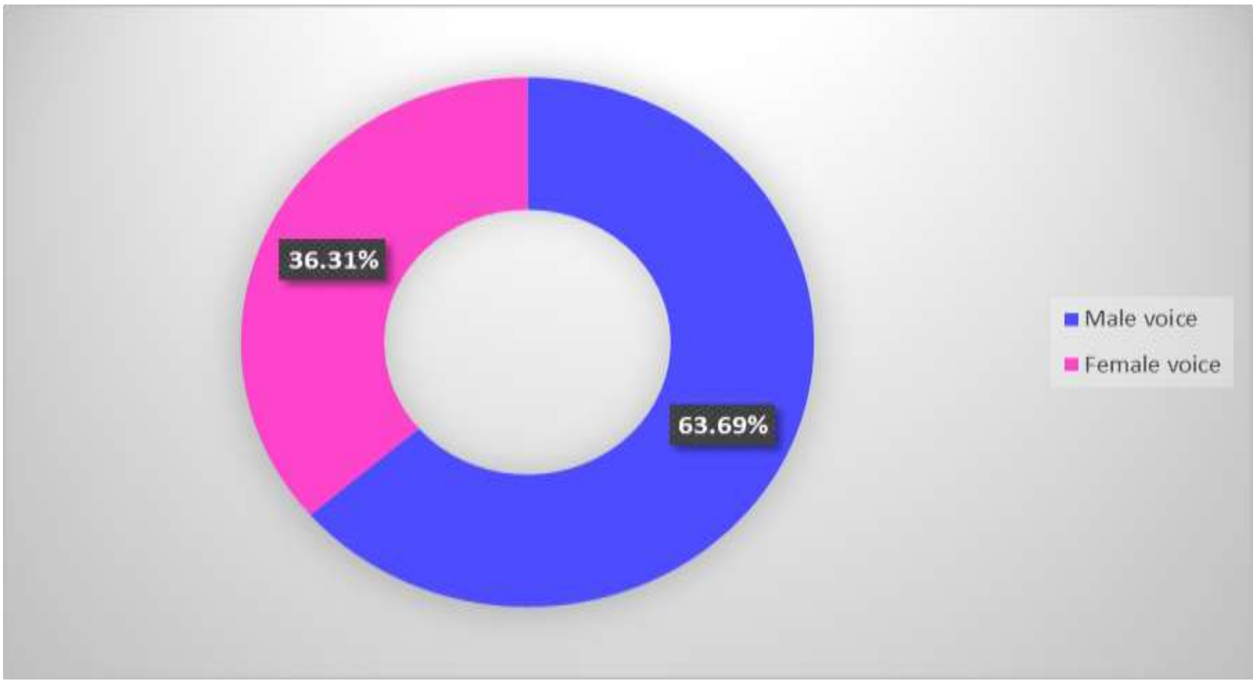

FIGURE 2: ratio of male and female speeches in the ArmSpeech corpus.

Overall male and female speeches ratio in Corpus (extension along with the first release) is 63.7% or 10 hours of male voice (1.75 times more than female speech) and 36.3% or 5.7 hours of female speech.

The extension contains 1 female and 3 male voices, but one of the male volunteers was the volunteer of voice for the last section of the first release. Hence force ArmSpeech's final release (first and second releases together) contains 13 male and 4 female voices.

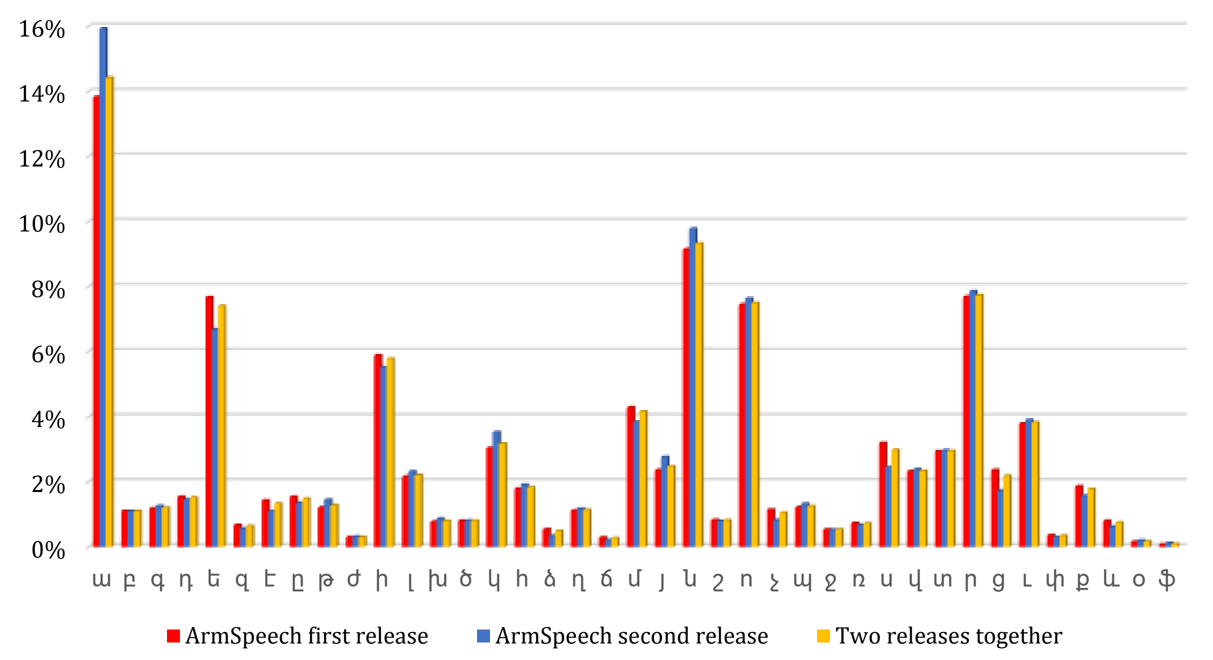

FIGURE 3: Distribution of the Armenian phonemes in the releases.

Statistics of the first release, extension and final corpus and other related information are presented in Table 1.

TABLE 1: Statistics of ArmSpeech corpus releases.

Specifications	ArmSpeech first release	ArmSpeech second release	Two releases together
Total duration	11:46:26	04:00:52	15:47:19
Minimum sample duration	0.72 seconds	0.62	0.62
Maximum sample duration	10.00 seconds	13.96	13.96
Mean sample duration	6.8 seconds	2.7	4.9
Total number of samples	6206	5378	11584
Total number of unique sentences (words or phrases)	6205	4838	11026
Total symbols	414685	160729	575414
Minimum number of symbols in samples	2	1	1
Maximum number of symbols in samples	144	135	144
Mean number of symbols in each sample	66.82	29.89	49.67
Total words	80632	26039	106671
Unique words	16847	9391	23062
Minimum number of words in samples	1	1	1
Maximum number of words in samples	31	19	31
Mean number of words in each sample	12.99	4.84	9.21

A full description of every section included in the second release is in Table 2.

TABLE 2: Description of ArmSpeech extended sections.

Section ID	Gender	Number of samples	Section duration in seconds	Age
14	Female	1844	~ 4586	49
15	Male	799	~ 2051	26
16	Male	2285	~ 7072	25
17	Male	450	~ 741	22

CONCLUSIONS

This paper introduced the second release of the ArmSpeech corpus. Along with the first release it can be used in text-to-speech, speech-to-text and spoken language identification systems research.

The extension contains 5378 annotated and high-quality audio clips totally of 4 hours. Along with the first release, the ArmSpeech corpus becomes 15.7 hours. Audio samples are mono channel, 16-bit WAV files with a 16000 Hz sampling rate and 256 kbps bit rate. The extension contains 3 male (63.69% of the speeches) and 1 female voice (36.31% of the speeches).

REFERENCES

[1] Wikipedia Contributors (2022). Speech corpus. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Speech_corpus.

[2] Wikipedia Contributors (2019). Natural language processing. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Natural_language_processing.

[3] Wikipedia Contributors (2019). Speech recognition. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Speech_recognition.

[4] Wikipedia Contributors (2019). Speech synthesis. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Speech_synthesis.

[5] Gundeep Singh, Sahil Sharma, Vijay Kumar, Manjit Kaur, Mohammed Baz and Mehedi Masud. Spoken Language Identification Using Deep Learning. Computational Intelligence and Neuroscience, Volume 2021, 21 September 2021.

[6] Wikipedia Contributors (2020). Acoustic model. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Acoustic_model.

[7] Baghdasaryan, V. H. (2022). ArmSpeech: Armenian Spoken Language Corpus. International Journal of Scientific Advances (IJSCIA), Volume 3| Issue 3: May-Jun 2022, Pages 454-459, URL: https://www.ijscia.com/wp-content/uploads/2022/06/Volume3-Issue3-May-Jun-No.283-454-459.pdf.

[8] Wikipedia Contributors (2019). Web scraping. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Web_scraping.

[9] Moaiad Khder. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing and its Applications, Volume 13, Issue 3, pages 145-168, December 2021. doi:10.15849/IJASCA.211128.11.

[10] Wikipedia Contributors (2022). Text normalization. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Text_normalization.

[11] Wikipedia Contributors (2021). WebRTC. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/WebRTC.

[12] Wikipedia Contributors (2021). Voice activity detection. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Voice_activity_detection.