International Journal of Scientific Advances
ISSN: 2708-7972
Volume: 3 | Issue: 3 | May - Jun 2022 Available Online: www.ijscia.com
DOI: 10.51542/ijscia.v3i3.25

ArmSpeech: Armenian Spoken Language Corpus

Varuzhan H. Baghdasaryan

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

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

ABSTRACT

The Armenian language is an independent branch of the Indo-European language family and the official language of the Republic of Armenia and the Republic of Artsakh. According to various reliable sources, an average of 3 million people in Armenia and 10-12 million people in the Armenian Diaspora use the Armenian language as their native language. The largest communities outside of Armenia are in the United States of America, Canada, the Russian Federation, the Islamic Republic of Iran, the French Republic, the Syrian Arab Republic and the Lebanese Republic. This paper presents the ArmSpeech speech corpus. ArmSpeech is a collection of annotated Armenian speech intended for natural language processing (NLP) technologies research and development. ArmSpeech is designed for speech-to-text and text-to-speech purposes but can be used in other domains also (e.g. language identification). Corpus contains 6206 high-quality audio samples: 11 hours 46 minutes and 26 seconds (11.77 hours) of annotated native Armenian speech of multiple speakers of any age, gender and accent. According to the research results, this is the most extensive Armenian speech corpus in the public domain for speech recognition, speech synthesis and spoken language identification systems.

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

INTRODUCTION

In our day communication between machines and humans is not only done by the graphical user interface (GUI) but also by natural human speech or by eyes gaze and gestures. For human-machine communication by natural human speech two mandatory and one optional technology are required: speech recognition (speech-to-text) [1] system that converts user human speech to text, speech synthesis (text-to-speech) [2] system that generates a meaningful human speech audio signal and spoken language identification [3] system that identifies currently spoken language. These three technologies require well-annotated speech corpora [4], which is a time-consuming process and requires knowledge of language grammar and lexicology. In the context of this paper, the words “annotation”, “annotate” or “annotated” will be used as the practice of appointing/tagging the corresponding labels (words or phrases that are in the audio clip speech) and other extra pieces of information to the audio clip.

It is preferred to train acoustic models [5] of speech-to-text and spoken language identification engines with multi-speaker corpora to extract all the features of language and increase the efficiency of the system when it is used by users of different genders, ages and accents.

Text-to-speech systems commonly use single-speaker corpora to train single-voice models, but many studies have shown [6,7] that combining multiple styles, timbres and training with multi-speaker corpora is the best way to get better quality, stable and multi-style synthetic speech.
Armspeech corpus is consist of 14 sections. Clips in the first 13 sections are from the public domain and free-to-use audiobooks of fiction (stories, novels, etc.). Clips of the final 14th section contain speeches about numerous real-life situations, including speeches about movies, sports, restaurants, as well as numbers, days of the week and months. Most of the sections are generated from fiction audiobooks because they contain both the most common words in the language and those words that are less common in ordinary everyday conversations and can only be found in fiction. In addition, readers of audiobooks sometimes change their tone and timbre of voice while reading speeches of different characters. This allows for adding more variety of speech to the corpus.

Although most of the clips do not contain leading and trailing silent parts, clips that contain silence at the beginning and at the end were not trimmed. All annotations (“audio clip – transcript” pairs) were additionally verified in order to avoid annotation errors, however, there is little chance of error.

The audio clips were released as mono-channel 16-bit files with a 16000 Hz sampling rate and 256 kbps bit rate. Audio samples were released in WAV (lossless compression) format because lossless audio file format like WAV is the best for sound quality and audio editing [8].

RELATED WORKS

There are many free or paid speech corpora for the most spoken languages available online. Some of them were thoroughly researched during the creation of ArmSpeech.
As mentioned in paper [9] the Common Voice corpus is a massively-multilingual collection of transcribed speech intended for speech technology research and development. Common Voice is designed for automatic speech recognition, language identification and speech synthesis systems. As of June 2022, the number of languages reaches 93. The data presented in this paper was collected and validated via Mozilla’s Common Voice initiative.

The recordings are later verified by other contributors using a simple voting system. The audio clips are released as mono-channel, 16-bit MPEG-3 files with a 48kHz sampling rate.

RyanSpeech [10] is a publicly available TTS corpus often noisy, recorded with multiple speakers, or lacks quality male speech data. RyanSpeech contains textual materials from real-world conversational settings. These materials contain over 9.84 hours of speech. All the audio files are recorded by a single professional male speaker with studio quality at a sampling rate of 44100 Hz.

RUSLAN [11] is an open Russian spoken language corpus for the text-to-speech task and contains 22200 audio samples with text annotations totally of 31 hours of speech of 23 years old male native Russian speaker. Audio samples were recorded in a quiet and noise-protected room using noise-reduction hardware and each sample was recorded separately with a sampling frequency of 44.1 kHz, 16-bit linear PCM and saved in WAV format. Leading and trailing silent parts were deleted from each audio sample. All text-audio pairs were additionally verified in order to avoid annotation errors.

Librispeech [12] is a CC-BY-licensed 1,000-hour standard large-scale speech dataset in the public domain collected from audiobooks of the Librivox project.

MLS [13] is a CC-BY-licensed 50000-hour speech dataset that is derived from the Librivox audiobooks. MLS is primarily a multilingual speech corpus, whereas the dataset focuses only on the English language but with a more diverse set of sources.

Earnings21 [14] is a 39 hour orthographically transcribed speech dataset of public companies earning calls created by expert transcriptionists and licensed under a CC-BY-SA license.

Gigaspeech [15] is a 10000-hour English speech dataset. Like the People's Speech, it uses forced alignment of existing audio against transcripts to create training data. However, it does not allow commercial usage because its sources may be copyrighted.

French Non-Native Corpus [16] consist of two parts. The first part includes common dialogue phrases in the tourism domain (e.g. hotel, restaurant, transport and others). The second part instead of dialogues contains sentences from tourism articles on the web. Seven native Chinese speakers and eight native Vietnamese (seven males and eight females) speakers have participated. The recording was done in a soundproof room, using a headset microphone, with a sampling frequency of 16 kHz. A supervisor was assigned to monitor and facilitate the recording of each speaker.

As can be seen, the common parameters of all these libraries are that they use the sampling frequency in the range of 16000-48000 Hz, and most of the audio clips were recorded in the professional rooms by professional speakers or donated by volunteers in high-quality audio formats.

DATA COLLECTION AND PRE-PROCESSING

ArmSpeech corpus sections creation process can be divided into two general stages:

- Free-to-use audiobook collection, data pre-processing and annotation.
- Recording and annotation of different pre-processed phrases or words collected from many spheres of life.

As mentioned in the introduction clips of the first 13 sections of ArmSpeech are from the public domain and free-to-use audiobooks of fiction. Audiobooks were manually collected from web resources in MP3 format. In parallel with this process, the corresponding books were also collected in text format. The audiobooks that are included in this corpus have high-quality audio and are recorded in noise-protected proof rooms. After collecting data, each of these audiobooks was divided into 10-seconds segments using the FFmpeg multimedia framework [17]. If the audiobook is available on the web in disassembled parts, those parts have been concatenated using FFmpeg before the segmentation process.

In the next step, each of the 10-second parts was checked to find out which clips contains complete and understandable phrases or words. Those 10-second clips which were not contained understandable phrases were marked as invalid and removed from the list.

Texts were normalized [18] according to 3 rules. Text data normalization steps are below:

- During the annotation all numbers and dates were manually replaced by their textual representation.
- Acronyms were not manually substituted with their expanded forms and because of this speech-to-text engines trained on ArmSpeech will have the ability to transcribe acronyms too.
- All symbols except for Armenian letters were automatically deleted including punctuation marks.

As mentioned above most of the clips do not contain leading and trailing silent parts. Clips that contain silence at the beginning and at the end were not trimmed.

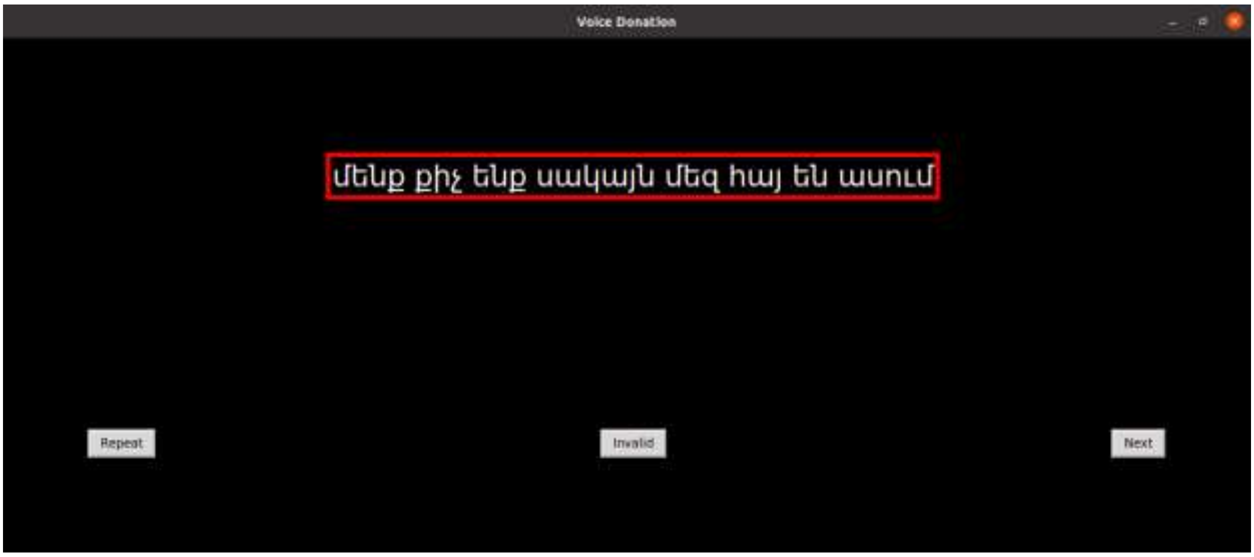
The last 14th section contains 2394 sentences (phrases about movies, sports, restaurants, as well as numbers, days of the week and months) which are recorded in a quiet and noise-protected proof room using a noise-reduction and echo-cancelling microphone and each sample has the same audio format and parameters as samples in other sections. Most parts of the data were scraped [19, 20, 21] from websites about daily news, lifestyle, culture, sport.

ANNOTATION

After data collection and pre-processing operations mentioned above, started the annotation process. The annotation of audio clips extracted from audiobooks (first 13 sections) was done simply by copying the corresponding text from the book text file to the list of "audio clip – transcript" pairs.

Annotation of the last 14th section was done using a small Python program which was created within the project frames. The interface of the application is in figure 1. This Python program simply shows a normalized sentence that must be annotated, and 3 buttons: "Repeat", "Invalid" and "Next".

The "Repeat" button gives the ability to record the current sentence (phrase or word) again, the "Invalid" button marks the current sentence as invalid and removes it from the sentences list. This button can be used when the sentence is not valid (e.g. grammatically incorrect words). After recording the "Next" button marks the current sentence as valid, includes transcript, recorde audio file size (in bytes) and recorded audio clip (in WAV format) in corpora and loads the next sentence for annotation.



մենք քիչ ենք սակայն մեզ հայ են ասում

FIGURE 1: the interface of the annotation application.

During the creation of ArmSpeech, the silence parts removing operation for the 14th section was done in the annotation stage.

The annotation program use voice activity detection technology to detect speech by simply distinguishing between silence and speech. By using voice activity detection technology recording starts when the speaker starts the speech and ends when the speaker ends the speech. This prevents including leading and trailing parts in the recorded audio clips.

Leading and trailing parts in the audio clips are usually removed to counter the fact that an automated speech recognition system will recognise any sound-phoneme combination which will make background noise (even if very low) and lead to an incorrect result.

STATISTICS OF THE CORPUS

The Armenian language is a relatively complex language. The Armenian alphabet has both lowercase and uppercase letters but as mentioned above during the text normalization stage all text data becomes lowercase therefore in the speech corpus are only lowercase letters of the Armenian alphabet.

The Armenian alphabet has 36 phonemes, which are written in 39 letters. The Armenian language has 6 vowels: “ա, է (ե), ը, ի, օ (ո), ու”, which are written in 8 letters and 30 consonants: “բ, գ, դ, զ, թ, ժ, լ, խ, ծ, կ, հ, ձ, ղ, ճ, մ, յ, ն, շ, չ, պ, ջ, ռ, ս, վ, տ, ր, ց, փ, ք, ֆ”.

Figure 2 shows the distribution of the Armenian phonemes in the ArmSpeech corpus.

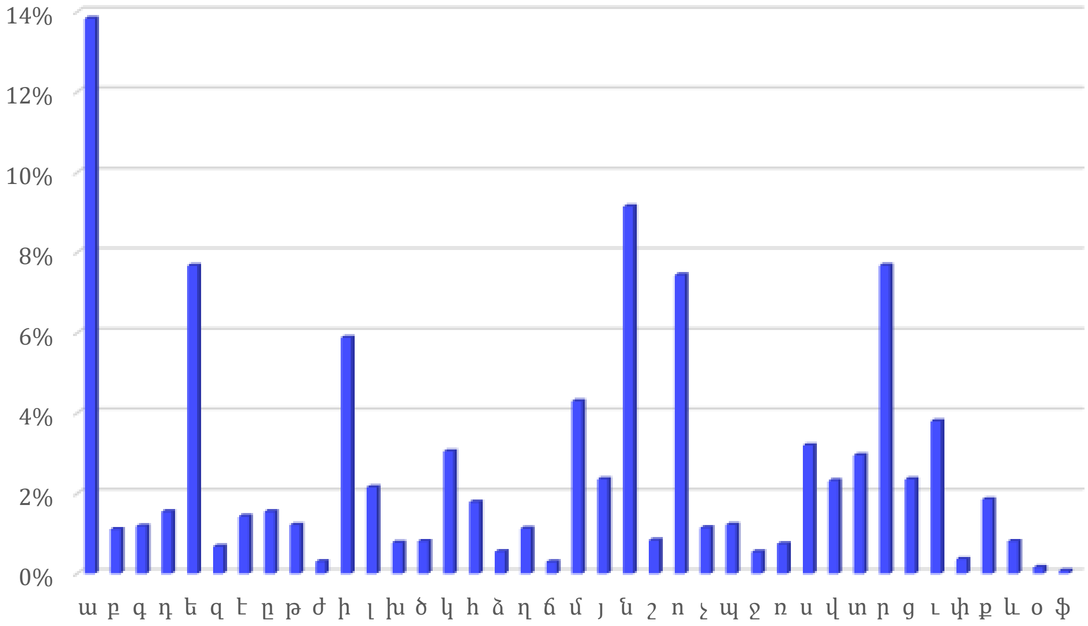


FIGURE 2: distribution of the Armenian phonemes in the corpus.

Corpus statistics and other related pieces of information are presented in table 1.

TABLE 1: ArmSpeech corpus statistics.

Total duration	**11:46:26**
Minimum sample duration	0.72 seconds
Maximum sample duration	10.00 seconds
Mean sample duration	6.8 seconds
Total number of samples	6206
Total number of unique sentences (words or phrases)	6205
Total symbols	414685
Minimum number of symbols in samples	2
Maximum number of symbols in samples	144
Mean number of symbols in each sample	66.82
Total words	80632
Unique words	16847
Minimum number of words in samples	1
Maximum number of words in samples	31
Mean number of words in each sample	12.99

11 of the sections contain a male voice, and 3 of the sections contain a female voice. As shown in figure 3 however male speeches (62.3720096% or over 7.3 hours) are 1.6 times more than female speeches (37.6279904% or over 4.4 hours).

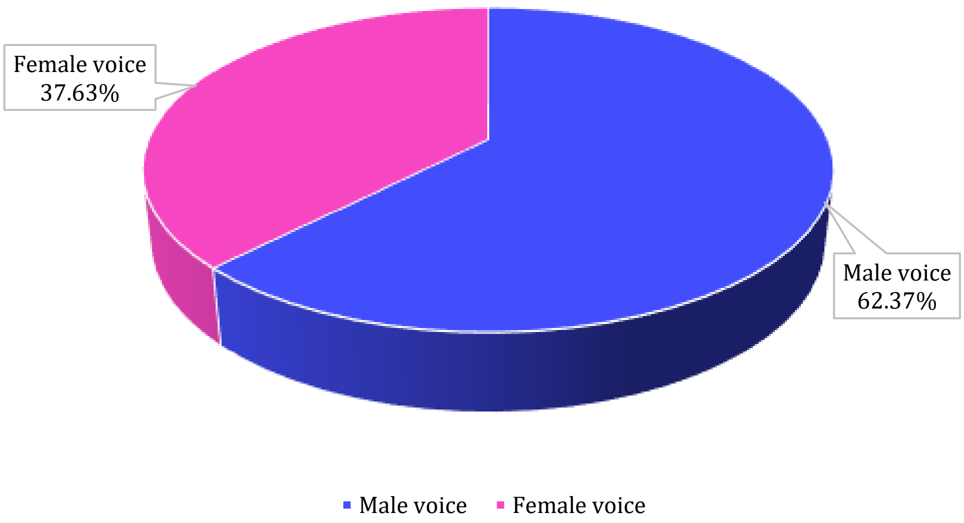


FIGURE 3: ratio of male and female speeches in the ArmSpeech corpus.

The audio clips are divided into train and validation sets. Division ratio is 80%-20%: corresponding 80% train set and 20% validation set. Train data information is stored in the "train.csv" file and validation data information is stored in the "validation.csv" file.

A full description of every section is in table 2.

TABLE 2: description of ArmSpeech corpus sections.

Section ID	**Gender**	**Number of samples**	**Section duration in seconds**	**Book title and author name**
0	Male	52	~ 510	Nar-Dos: "I and She"
1	Female	970	~ 9682	Raffi: "The Fool"
2	Male	209	~ 2087	Garegin Nzhdeh: "Sons struggle against fathers" (also includes Garegin Nzhdeh's autobiography and the article "Will to die")
3	Male	71	~ 709	Alexander Shirvanzade: "The captive of the idea"
4	Male	58	~ 579	Arthur Conan Doyle: "Sherlock Holmes: The Boscombe Valley Mystery"

Section ID	Gender	Number of samples	Section duration in seconds	Book title and author name
5	Male	553	~ 5521	Daniel Defoe: “Robinson Crusoe”
6	Female	207	~ 2061	Antoine de Saint-Exupéry: “The Little Prince”
7	Male	298	~ 2976	George Orwell: “Animal Farm”
8	Male	389	~ 3885	F. Scott Fitzgerald: “The Great Gatsby”
9	Female	421	~ 4204	Albert Camus: “The Stranger”
10	Male	68	~ 679	Arthur Conan Doyle: “Sherlock Holmes: The Adventure of the Copper Beeches”
11	Male	81	~ 809	Arthur Conan Doyle: “Sherlock Holmes: The Red-Headed League”
12	Male	435	~ 4344	Arthur Conan Doyle: “Sherlock Holmes: The Hound of the Baskervilles” (chapters: 1, 2, 3, 4, 5, 6, 10, 11, 12, 13, 14, 15)
13	Male	2394	~ 4334	Phrases about movies, sports, restaurants, as well as numbers, days of the week and months.

SPEECH CORPUS STRUCTURE

For storing validated audio clips and other extra pieces of information, has been chosen the CSV [22] file format. CSV files are designed for keeping comma-separated values of records. The released version of ArmSpeech contains 3 CSV files:

- Validated.csv – contains all validated audio clips data information.
- Train.csv – contains train set information.
- Validation.csv – contains validation set information.

The CSV files consist of 3 columns: “wav_filename”, “wav_filesize” and “transcript”. “wav_filename” column keeps the audio clip name (the relative path), “wav_filesize” column keeps the corresponding sample size given in bytes and “transcript” column keeps corresponding transcripts. This structure of corpora is the same as the structure of Common Voice’s corpora after bringing it into a form that DeepSpeech [23] understands.

CONCLUSIONS

This paper introduced the ArmSpeech corpus, which is a multi-speaker dataset designed for speech-to-text, text-to-speech and spoken language identification systems. The speech corpus contains 6206 annotated and high-quality audio clips totally of 11.77 hours. Audio samples are mono-channel, 16-bit WAV files with a 16000 Hz sampling rate and 256 kbps bit rate. Corpus contains 11 male (62.37% of the speeches) and 3 female voices (37.63% of the speeches). Audio clips of the first 13 sections were collected from the public domain and free-to-use audiobooks and then annotated with double validation steps. Although the corpus mainly consists of audiobooks, it also contains speeches about numerous real-life situations. The last 14th section of the ArmSpeech corpus contains phrases about movies, sports, restaurants, as well as numbers, days of the week and months recorded in a quiet and noise-protected proof room using a noise-reduction and echo-cancelling microphone.

According to research, ArmSpeech is the first large, high-quality multi-speech Armenian corpus that is corpus freely available for download. ArmSpeech can be the best choice for research in the natural language processing tasks using the Armenian language.

FURTHER WORK

According to the results of the research, in terms of corpora, speech-to-text, text-to-speech and language identification systems, the Armenian language is one of the low-resource languages. ArmSpeech speech corpus gives a huge ability to train, test and evaluate systems mentioned below and increase the resources.

In near future, it is planned to increase the size of the corpus by releasing the second version, which will include annotated speech collected from the content of the Armenian news websites.

It is also planned to test and evaluate the corpus by creating Armenian language acoustic and language models [24].

REFERENCES

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

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

[3] 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.

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

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

[6] Qicong Xie, Tao Li, Xinsheng Wang, Zhichao Wang, Lei Xie, Guoqiao Yu, Guanglu Wan. Multi-speaker Multi-style Text-to-speech Synthesis with Single-speaker Single-style Training Data Scenarios. arXiv:2112.12743v1, 23 December 2021.

[7] Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, Nobuyuki Nishizawa. Training Multi-Speaker Neural Text-to-Speech Systems using Speaker-Imbalanced Speech Corpor. arXiv:1904.00771v2, 7 April 2019.

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

[9] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. arXiv:1912.06670v2, 5 March 2020.

[10] Rohola Zandie, Mohammad H. Mahoor, Julia Madsen, Eshrat S. Emamian. RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis. arXiv:2106.08468v1, 15 June 2021.

[11] Lenar Gabdrakhmanov, Rustem Garaev, Evgenii Razinkov. RUSLAN: Russian Spoken Language Corpus for Speech Synthesis. arXiv:1906.11645v1, 26 June 2019.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An ASR Corpus Based on Public Domain Audio Books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210, 19-24 April 2015. doi: 10.1109/ICASSP.2015.7178964.

[13] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert. MLS: A Large-Scale Multilingual Dataset for Speech Research. Interspeech 2020, October 2020. doi: 10.21437/Interspeech.2020-2826.

[14] M. D. Rio, N. Delworth, R. Westerman, M. Huang, N. Bhandari, J. Palakapilly, Q. McNamara, J. Dong, P. Zelasko, and M. Jette. Earnings-21: A Practical Benchmark for ASR in the Wild. Interspeech 2021, 30 August - 3 September 2021.

[15] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Kudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan. GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio. Interspeech 2021, 30 August - 3 September 2021.

[16] Tien-Ping Tan, Laurent Besacier. A French Non-Native Corpus for Automatic Speech Recognition. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), May 2006.

[17] H. Sumesh Singha, Dr. Bhuvana J. A Study on FFmpeg Multimedia Framework. International Journal of Trend in Scientific Research and Development (IJTSRD), Volume 5, Issue 4, June 2021.

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

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

[20] 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.

[21] Lusiana Citra Dewi, Meiliana, Alvin Chandra. Social Media Web Scraping using Social Media Developers API and Regex. The 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019): Enabling Collaboration to Escalate Impact of Research Results for Society, 12–13 September 2019, Procedia Computer Science, Volume 157, 2019, pages 444-449. doi: https://doi.org/10.1016/j.procs.2019.08.237

[22] Wikipedia Contributors (2020). Comma-separated values. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Comma-separated_values.

[23] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng. Deep Speech: Scaling up end-to-end speech recognition. arXiv:1412.5567v2, 19 December 2014.

[24] Wikipedia Contributors (2019). Language model. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Language_model.