Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 4 (22)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

UDC 004.9
DOI: https://doi.org/10.30837/ITSSI.2022.22.005

O. BARKOVSKA

RESEARCH INTO SPEECH-TO-TEXT TRANFROMATION MODULE IN THE PROPOSED MODEL OF A SPEAKER'S AUTOMATIC SPEECH ANNOTATION

The **subject** matter of the article is the module for converting the speaker's speech into text in the proposed model of automatic annotation of the speaker's speech, which has become more and more popular in Ukraine in the last two years, due to the active transition to an online form of communication and education as well as conducting workshops, interviews and discussing urgent issues. Furthermore, the users of personal educational platforms are not always able to join online meetings on time due to various reasons (one example can be a blackout), which explains the need to save the speakers' presentations in the form of audio files. The **goal** of the work is to elimination of false or corrupt data in the process of converting the audio sequence into the relevant text for further semantic analysis. To achieve the goal, the following **tasks** were solved: a generalized model of incoming audio data summarization was proposed; the existing STT models (for turning audio data into text) were analyzed; the ability of the STT module to operate in Ukrainian was studied; STT module efficiency and timing for English and Ukrainian-based STT module operation were evaluated. The proposed model of the speaker's speech automatic annotation has two major functional modules: speech-to-text (STT) i summarization module (SUM). For the STT module, the following **models** of linguistic text analysis have been researched and improved: for English it is wav2vec2-xls-r-1bz, and for Ukrainian it is Ukrainian STT model (wav2vec2-xls-r-1b-uk-with-lm.Artificial neural networks were used as a mathematical apparatus in the models under consideration. The following **results** were obtained: demonstrates the reduction of the word error level descriptor by almost 1.5 times, which influences the quality of word recognition from the audio and may potentially lead to obtaining higher-quality output text data. In order to estimate the timing for STT module operation, three English and Ukrainian audio recordings of various length (5s, ~60s and ~240s) were analyzed. The results demonstrated an obvious trend for accelerated obtaining of the output file through the application of the computational power of NVIDIA Tesla T4 graphic accelerator for the longest recording. **Conclusions**: the use of a deep neural network at the stage of noise reduction in the input file is justified, as it provides an increase in the WER metric by almost 25%, and an increase in the computing power of the graphics processor and the number of stream processors provide acceleration only for large input audio files. The following research of the author is focused on the study of the methods of the obtained text summarization module efficiency.

Keywords: STT; text; processing; summary; audiofile; model; neural networks.

Introduction

The period of the pandemic and the war conflict has urged the development and expansion of various digital educational platforms feature set [1, 2]. These information spaces are necessary at the various education levels in different countries all over the world – from primary schools to higher education institutions as well as educational courses in various business spheres, thus enabling to provide students with learning materials, communication with teachers as well as remote knowledge level assessment. Therefore, virtual interaction of the distributed user community is established [3–4]. The amount of interactive features of digital education platforms is constantly expanded, thus, giving more opportunities to students (students of higher education institutions, postgraduate students, course participants etc.)

One example of the expanded feature set is access to audio files available for listening, but not aimed for text file production. Thus, provision of this function requires transformation of audio files into text preserving only valuable and relevant information [5].

The given work proposes a model of speech text annotation formation. One of the research lines is speech processing and transformation of audio files into text at the same time preserving only valuable and relevant information. The topicality and, simultaneously, the difficulty lies in the fact that spontaneous speaking is unstructured, does not resemble the written materials, and includes bits of information, which is repeated or corrected [6]. This requires adaptation and improvement of the existing STT approaches to certain conditions – peculiarities of the speaker's speech style, sensor properties, and system requirements as for the end result.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2022. No. 4 (22)

Analysis of last achievements and publications

Speaking is the simplest form of communication; there exist certain problems with speech recognition such as speech fluidity, pronunciation, words confusion, speech impediment problems. These must be solved during speech processing [7]. Moreover, environment peculiarities add up in the process of audio materials recording.

The current speech recognition systems have undergone a long process of development from their old analogs. They can recognize the speech of several speakers and use an enormous vocabulary in numerous languages [8–9].

The first experiments date back to the 1970s, but the developments in the sphere of parallel and distributed computing architectures, big data and artificial intelligence in the last years have given a great impetus to improve this technology and, thus, its reliability [10–11]. Compared to the past, the accuracy of transcription has actually improved to such a level that, on condition of a clear and clearly defined acoustic source, the accuracy level may well exceed 99%.

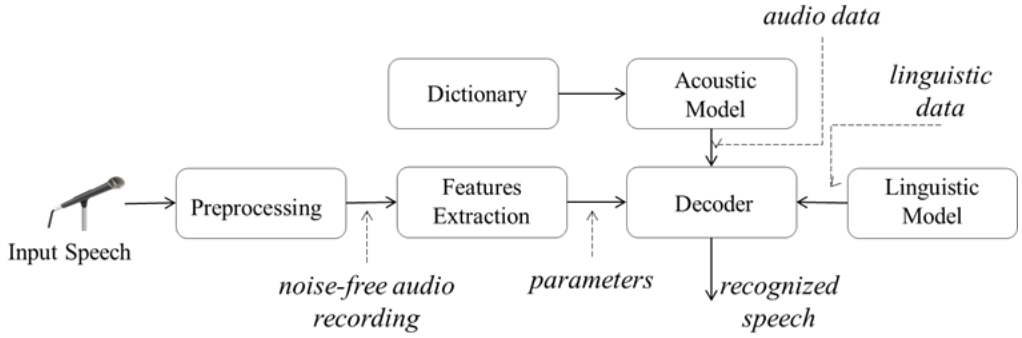

Fig. 1. Typical Speech-to-Text Scheme

The automatic transcription system core is automatic speech recognition, which combines the acoustic and the speech components (fig. 1). The acoustic component is responsible for the transformation of audio files into a sequence of tiny acoustic units. The "analog sound", i.e. vibrations produced during the speech, is transformed into digital signals, which can be scanned by software. Then, acoustic units are associated with the existing "phonemes", namely, the sounds, used in a certain language to form meaningful expressions. So, the linguistic component is responsible for the transformation of sequences of acoustic units into words, sentences and paragraphs. The linguistic component analyses all the previous words and their correlation in order to evaluate the possibility of applying a certain word in further speech. In technical terms, they are called "Hidden Markov Models" and are widely used in all speech recognition software [12]. Both components must be correctly "taught" to understand a certain language: the acoustic and the linguistic components are equally critical for transcription accuracy [13]. Figure 1 shows a flow diagram of a typical speech-to-text (STT) transformation system.

Certain advanced technological solutions by various technological companies currently exist. However, every solution has its own advantages and disadvantages, provided further (table 1). The following most common disadvantages are inherent for the program solutions under study:

- high cost;
- limited language support capabilities;
- lack of possibility to modify solutions.

When transcribing speech documents, the speech is divided into spontaneous and prepared. A much higher accuracy of recognition can be achieved when transcribing the speech, read from the text, for example, a newsreader's speech during a news broadcast. The capability of automatic spontaneous speech recognition is currently limited due to the lack of the structure and the presence of repetitions and corrections. Transcription of audio data into text is made after the transformation of a physical signal into an electrical signal via analog-to-digital converters. The widespread STT conversion methods include:

- hidden Markov models (HMM)
- deep neural networks (recurrent neural networks, convolution neural networks).

The selected criteria for the existing solution methods comparison were:

- the amount of data necessary for learning;
- the speed of learning;
- recognition accuracy.

Dynamic time warping (DTW) as a relevant estimation method to determine the similarity of

Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 4 (22)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

two sequences different with regard to time and speed was earlier popular in speech recognition, but now has been replaced by more effective methods based on hidden Markov models. DTW was applied to analyse video, audio and graphic files as well as any data capable of being converted into linear representation [14]. This can be exemplified by determination of similarities in walking patterns if a person walked faster in one video and more slowly in another one.

Automatic speech recognition is a well-known task for working with variable frequency of speech. If formulated differently, sequences (e.g. time series) are distorted nonlinearly. Therefore, the results obtained applying DTW method are not satisfactory.

Table 1. *Comparison of the Existing Solutions*

Criteria	Program solutions		
	Amazon Transcribe	**Dragon Anywhere**	**QuillBot**
Support of voice conversion into text	supported	supported	no
Support of summarization feature	no	no	supported
Feature set, operating with Ukrainian	supported	no	no
Feature set, operating with English	supported	supported	supported
Operation mode variation	no	no	no

Phonemes (the simplest units of human speech) probabilities distribution in the audible alarm segment, with a certain probability, enables to distinguish hidden Markov models. This is necessary for further reproduction of what was said by the sound. Taking into account the fact that the same phoneme may sound differently (e.g. depending on the accent), the choice of the feature (phoneme) probability distribution function is stipulated by the possibility to make allowance for and summarize several probability distributions (namely, take into account different sounding of the same phoneme). Gaussian Mixture Model (GMM) best meets the abovementioned requirement. GMM-HMM model is so successful that any new method can hardly outperform it for acoustic modelling. Despite all the advantages, GMM has a substantial drawback – inefficiency for modelling data, with lie on or near the nonlinear variety in the data space.

An alternative way of speech recognition is deep neural network (DNN), which has a lot of hidden layers and learns by means of new methods, outperforming GMM, sometimes by far, in various speech recognition tests [15].

Due to the abovementioned drawbacks of HMM-based models along with the development of deep learning technologies, end-to-end LVCSR is used in more and more solutions. The end-to-end model is the system, which directly converts an input acoustic sequence into a word sequence or another grapheme. Its functional structure is presented in figure 2.

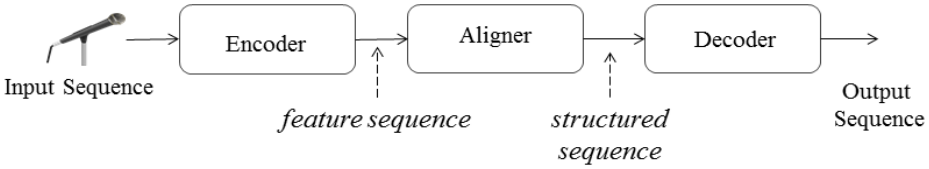

Fig. 2. Functional Structure of End-to-End Model

The majority of end-to-end speech recognition models include the following parts:

– encoder, which maps a sequence of speech input into a sequence of features;

– aligner, which implements alignment between the objects "sequence" and "language";

– decoder, which decodes the end result of the identification.

This distribution does not always take place because the end-to-end model itself is a seamless structure and it is usually difficult to define which subtask is performed by which part.

Compared to an HMM-based model, an LVCSR model has the following advantages:

– several modules are combined into one network for collaborative learning. i.e. there is no need to represent the middle states;

– there is no data alignment problem, in particular for language recognition because the "subtle approach" to alignment is used, in which every audio frame

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2022. No. 4 (22)

corresponds to all the possible states with certain probability distribution.

The aim of the work is to reduce false or corrupt data occurrence in the process of conversion of a sound series into a relevant text for further semantic analysis for English and Ukrainian input audio files.

In order to achieve the set aim, the following tasks must be solved:

– development of a generalized text summarization model for input audio data;

– review and analysis of the existing STT models (Speech-to-text models);

– provision of STT module processing of the Ukrainian language;

– quality assessment of STT modules working with Ukrainian and English;

– STT module operation timing evaluation;

– analysis of the obtained results.

Audio-to-text conversion under study is one of the modules of the hybrid model proposed in this work, which enables to recognize the speaker's speech, convert the available audio data into text and summarize the text obtained after input audio materials conversion, which is the final stage, preserving only the informative part of the lecture.

An important feature of the solution proposed by the author might be distributed data processing and scalability, which enable to apply the given approach to big data processing. The task of voice and text processing requires considerable resources, therefore, concurrent GPGPU processing paradigm is applied in order to optimize and accelerate the given task.

Materials and methods

Transcription of speech documents such as public speeches, oral project presentations, lectures and TV news are one of the basic applications of automatic speech recognition. Although speech is the most natural and successful form of human interaction, it is hard to quickly evaluate, obtain and reuse text documents, which are simply written as acoustical signals.

The given work researches the use of several speech-to-text models as components of the proposed speaker's speech annotation model. In order to fulfill the given task, a decision was made to develop modular microservice architecture, which would provide for semantic models change with minor impact on the whole complex [15].

Speech-to-text module receives a WAV sound record with the frequency of 16 kHz as input as this is the limit to audio series conversions and it requires high-quality input. Several audio series processing models have an urgent problem of obtaining an audio without ambient noise [16 – 17], which is not implemented in some solutions. The given function is not implemented in the model selected for research, either. Therefore, a decision was made to approach the task of audio series cleanup via a deep neural network, described in [18].

Text summarization module takes the result of speech recognition (ASR) module as input in the form of a JSON object. The given research also puts forward the idea to develop a service for input text filtering by means of marker words. Text filtering takes place at the stage of JSON object transition from the ASR module to the text summarization module.

Figure 3 shows the proposed automatic speaker's speech annotation model. The main functional modules of the proposed model are: speech-to-text (STT) and summarization modules (SUM). The main focus of the work lies on the STT module and the main task is automatic speech recognition (ASR).

Two models were tested for the given module: wav2vec2-xls-r-1b₃ for English, which was an open source model by Facebook and available in the public domain; Ukrainian STT model (wav2vec2-xls-r-1b-uk-with-lm) for Ukrainian, which is a revised version of Facebook solution.

Traditional speech recognition models are primarily trained on transcribed and annotated speech audios; they require annotated big data, available for only a few languages.

In this case, model training differs – it is asked to predict what the speaker tells further through comparison of several options. The given approach studies a set of language units, which are shorter than phonemes, to describe the audio sequence of the speech. For the reason this set is finite, the model cannot represent all variations, such as background noise. On the contrary, units prompt the model to focus on the most important factors for speech audio representations.

The wav2vec2 model first processes the raw form of the speech audio signal using a multilayer convolutional neural network to produce latent sound representations of 25 ms each. This model studies the basic language units used to solve a self-monitored task. The model is trained to predict the correct language unit for the masked parts of the audio, along with learning what the language units should be. With only 10 minutes of transcribed speech

Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 4 (22)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

and 53 thousand hours of unlabeled speech, wav2vec 2.0 can simulate speech recognition with a word error rate (WER) of 8.6 percent for noisy language and 5.2 percent for pure language on the standard LibriSpeech test.

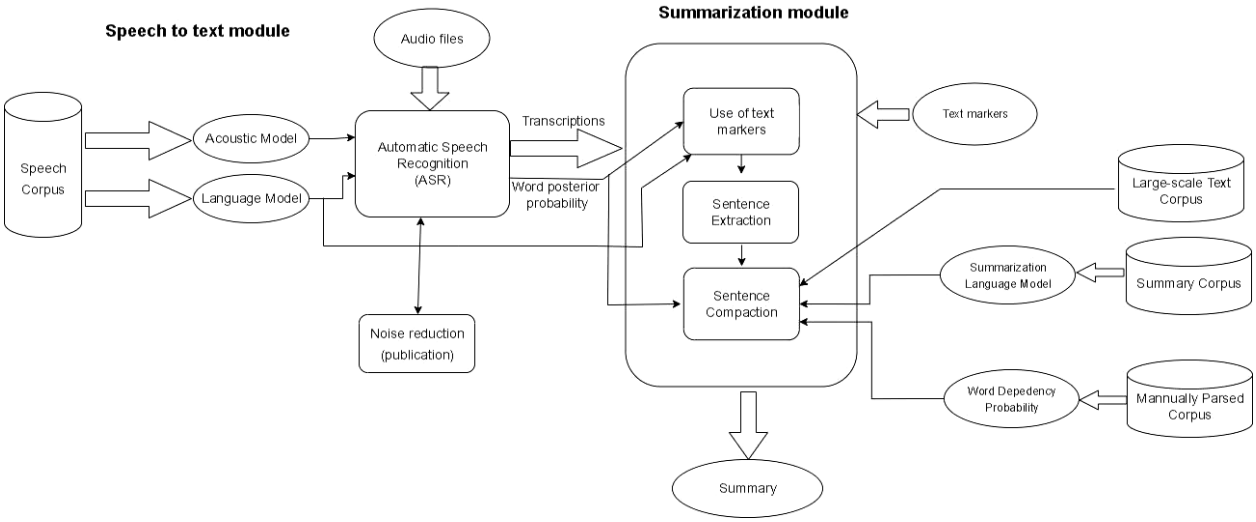

Fig. 3. The Proposed Automatic Speaker's Speech Annotation Model

For wav2vec, an architecture was developed, which consists of two-layered convolutional neural networks laid over each other. The encoder network maps raw audio input into a representation, in which each vector covers about 30 milliseconds (ms) of speech. The context network uses these vectors to create its own representations covering a larger span up to the second.

The number of neural layers in the extractor module is 7. The mozilla-foundation/common_voice_7_0 was used as the dataset to train the model. Common Voice dataset consists of unique wav recordings and corresponding text files. The major part of the 13,905 recorded hours in the dataset also contain demographic metadata, such as age, gender, and accent, which can help improve the accuracy of speech recognition mechanisms.

The dataset currently consists of 11,192 verified hours in 76 languages, but more voices and more languages are added all the time. An additional linguistic model was used for the model, which supports the Ukrainian language.

The results of the Common Voice 7 (WER) test evaluation without and with the additional Ukrainian linguistic model are shown in table 2.

On the basis of the obtained results, shown in table 2, a conclusion can be made about the reduction of the word error rate when using the additional Ukrainian linguistic model to recognize audio files, which are recorded in the Ukrainian language. This affects the quality of word recognition in the audio and can potentially provide for a better quality of output text.

Table 2. *WER Results for Model, Which Supports Ukrainian*

With linguistic model	Without linguistic model
14.62	21.52

In the process of preparation for the wav2vec2 model test, additional training of Ukrainian STT model was conducted; the results of training are presented below in table 3.

In contrast to conventional methods of minimum square error (MMSE)-based noise reduction, the proposed supervised speech enhancement method by means of the mapping function search between noisy and pure speech signals is performed on the basis of deep neural networks (DNNs). In order to be able to deal with a wide range of additive noise in real-life situations, a large training set covering many possible combinations of speech and noise types was initially developed.

The given DNN model was originally trained on 100 hours of noise speech data with 104 noise types. To improve the generalization ability of DNN under noise mismatch conditions, 3 hidden layers and 2048 hidden units for each hidden layer were used.

The experiment results demonstrate that the proposed framework can achieve significant improvements over the conventional MMSE-based technique. It is also worth noting that the proposed DNN approach can remove transient noise, which is difficult to process successfully. Furthermore, the obtained DNN model trained on artificially synthesized data is also effective for handling

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2022. No. 4 (22)

noisy speech data recorded in real-world scenarios, without creating the annoying musical artifact typical for conventional enhancement techniques.

In table 4, a WER-metric-based comparison is presented with respect to its performance based on the rounded-up averages of the 3 wav format audio tracks.

In the case of using the Ukrainian language, this difference is important because the WER, being in the normal range when lecture speech is recognized, is 20–30.

In order to evaluate the STT module performance when using additional ukrainian linguistic model, 3 audio recordings in English and Ukrainian with different length were analyzed. Recordings classification is: short (5s), medium (~60s), long (~240s). The results of the time spent to process the audio signal in Ukrainian and English in the cloud solutions [19] and on the personal computer are shown in figure 4.

The experiment was conducted using computers with different performance. The following hardware was used as an available computer on a personal computer – a central processor Intel Core i7-9750H (2.6 – 4.5 GHz), a graphic processor NVIDIA GeForce GTX 1650 Mobile. The hardware characteristics of the remote cloud solution are as follows – Intel Xeon 2.30GHz CPU, NVIDIA Tesla T4 GPU.

With reference to the obtained results, the trend for accelerating the longest recording under study is obvious. Word error rate (WER) was also measured for cleaned audio, which is the same for the both video cards, but different for the languages under study. The results of the measurements can be seen in table 5.

With regard to the obtained results, there is an upward trend for the level of errors in words along with data amount increase. This may be due to insufficient training of the model for the Ukrainian language and lead to the loss of the logical meaning of the record.

Table 3. *Results of wav2vec2 Model Training*

Training Loss	Epoch	Step	Validation Loss	Wer	Cer
1.2815	7.93	500	0.3536	0.4753	0.1009
1.0869	15.86	1000	0.2317	0.3111	0.0614
0.9984	23.8	1500	0.2022	0.2676	0.0521
0.975	31.74	2000	0.1948	0.2469	0.0487
0.8868	47.61	3000	0.1903	0.2257	0.0439
0.8424	55.55	3500	0.1786	0.2206	0.0423
0.8126	63.49	4000	0.1849	0.2160	0.0416
0.7901	71.42	4500	0.1869	0.2138	0.0413
0.7671	79.36	5000	0.1855	0.2075	0.0394
0.7467	87.3	5500	0.1884	0.2049	0.0389
0.731	95.24	6000	0.1877	0.2060	0.0387

Table 4. *Ambient Noise Cleanup Efficiency*

Language	**Ukrainian**	**English**
Audio series, cleared from noise, WER	29	3.9
Audio series, not cleared from noise, WER	38.7	5.2
Difference, %	25%	21%

Table 5. *Word Error Rate (WER) Measurement Results for Cleared Sound Series in Ukrainian and English*

Audio Series Length	**Ukrainian, WER**	**English, WER**
Short	24.6	3.4
Medium-length	28.4	3.7
Long	33.9	4.5

Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 4 (22)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

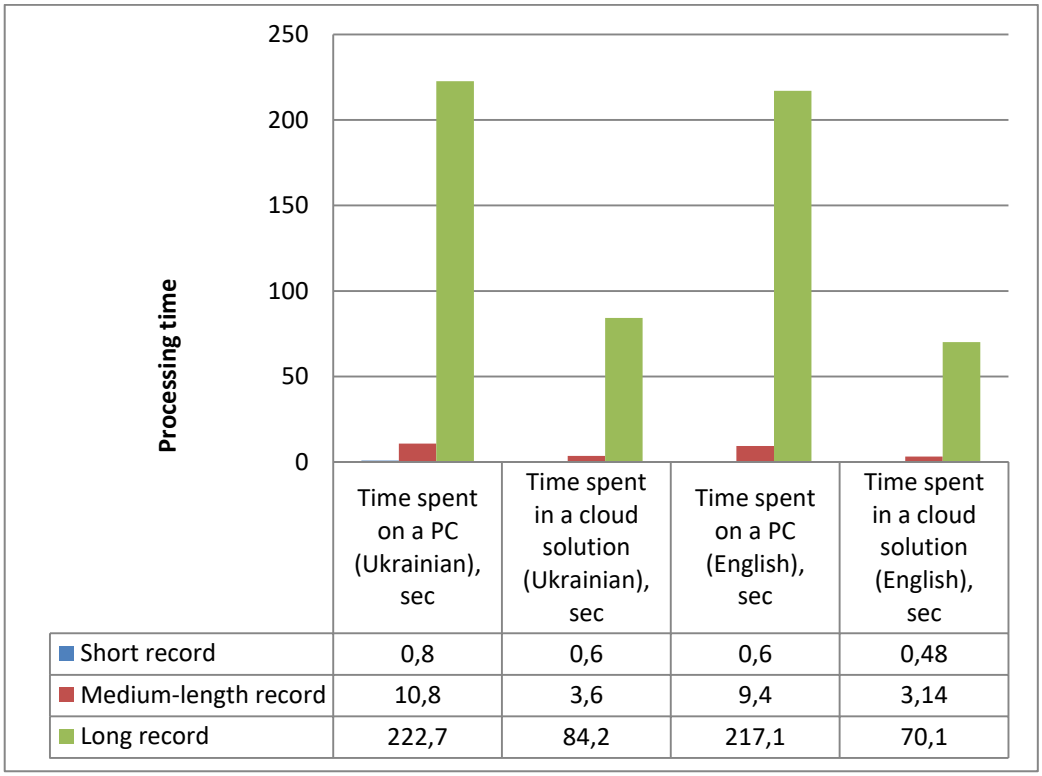

Fig. 4. Bar Chart of Time Spent on 5s, 60s and 240s-Long Audio Recordings

Conclusion

In order to reduce the appearance of false or distorted data during the conversion of a sound series into relevant text for further semantic analysis for input audio files in Ukrainian and English, a generalized model of incoming audio data summarization was proposed; the existing STT models (for turning audio data into text) were analyzed; the ability of the STT module to operate in Ukrainian was studied; STT module efficiency and timing for English and Ukrainian-based STT module operation were evaluated. The proposed model of the speaker's speech automatic annotation has two major functional modules: speech-to-text (STT) i summarization module (SUM). Two models were studied and improved for the STT module. For the English language, this is wav2vec2-xls-r-1bз and for the Ukrainian language, this is Ukrainian STT model (wav2vec2-xls-r-1b-uk-with-lm). wav2vec2, improved using the Ukrainian linguistic model, demonstrates the reduction of the word error level descriptor by almost 1,5 times, which influences the quality of word recognition from the audio and may potentially lead to obtaining higher-quality output text data. The application of a deep neural network at the input file noise suppression stage is also well-grounded as it provides for the increase in WER metric by nearly 25%. In order to estimate the timing for STT module operation, three English and Ukrainian audio recordings of various length (5s, ~60s & ~240s) were analyzed. The results demonstrated an obvious trend for accelerated obtaining of the output file through the application of the computational power of NVIDIA Tesla T4 graphic accelerator for the longest recording. The following research of the author is focused on the study of the methods of the obtained text summarization module efficiency.

References

1. Liu, J., Wang, H. (2021), "An Analysis of the Educational Function of Network Platform from the Perspective of Home-School Interaction in Universities in the New Era", *2021 IEEE International Conference on Educational Technology (ICET)*, 2021, P. 112–116. DOI: https://doi.org/10.1109/ICET52293.2021.9563158
2. Ponomarova, H., Kharkivska, A., Petrichenko, L., Shaparenko, K., Aleksandrova, O., Beskorsa, V. (2021), "Distance Education In Ukraine In The Context Of Modern Challenges: An Overview Of Platforms", *International Journal of Computer Science & Network Security*, 21 (5), P. 39–42. DOI: https://doi.org/10.22937/IJCSNS.2021.21.5.7

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2022. No. 4 (22)*

3. Berrío-Quispe, M. L., Chávez-Bellido, D. E., González-Díaz, R. R. (2021), "Use of educational platforms and student academic stress during COVID-19," *2021 16th Conference on Information Systems and Technologies (CISTI)*, P. 1–5. https://doi.org/10.23919/CISTI52073.2021.9476308
4. Malieieva, J., Kosenko, V., Malyeyeva, O., & Svetlichnyj, D. (2019), "Creation of collaborative development environment in the system of distance learning", *Innovative Technologies and Scientific Solutions for Industries*, 2 (8), P. 62–71. DOI: https://doi.org/10.30837/2522-9818.2019.8.062
5. Dong, Q., Ye, R., Wang, M., Zhou, H., Xu, S., Xu, B., & Li, L. (2021), "Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 14, P. 12749–12759.
6. Gao, Jianqing, Wan, Genshun, Wu, Kui and Fu, Zhonghua (2022), "Review of the application of intelligent speech technology in education", *Journal of China Computer-Assisted Language Learning*, Vol. 2, No. 1, P. 165–178. DOI: https://doi.org/10.1515/jccall-2022-0004
7. Liu, J., Xiang, X. (2017) "Review of the anti-noise method in the speech recognition technology," *12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, P. 1391–1394. DOI: https://doi.org/10.1109/ICIEA.2017.8283056
8. Juang, Biing-Hwang, and Lawrence, R. Rabiner (2005), *Automatic speech recognition–a brief history of the technology development*, Georgia Institute of Technology, Atlanta Rutgers University and the University of California, Santa Barbara 1, 67 p.
9. Potamianos, G. (2009), "Audio-visual automatic speech recognition and related bimodal speech technologies: A review of the state-of-the-art and open problems," *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, P. 22–22. DOI: https://doi.org/10.1109/ASRU.2009.5373530
10. Georgescu, A. L., Pappalardo, A., Cucu, H., & Blott, M. (2021), "Performance vs. hardware requirements in state-of-the-art automatic speech recognition", *EURASIP Journal on Audio, Speech, and Music Processing*, No.1, P. 1–30.
11. Mohammed, A., Sunar, M. S., & hj Salam, M. S., (2021), "Speech recognition toolkits: a review", *The 2ndNational Conference for Ummah Network 2021 (INTER-UMMAH 2021)*, No. 2, P. 250–255.
12. Kumar, T., Mahrishi, M., & Meena, G. (2022), "A comprehensive review of recent automatic speech summarization and keyword identification techniques", *Artificial Intelligence in Industrial Applications*, P. 111–126.
13. Kim, C. *et al*. (2019), "End-to-End Training of a Large Vocabulary End-to-End Speech Recognition System," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, P. 562–569, DOI: https://doi.org/10.1109/ASRU46091.2019.9003976
14. Ping, L. (2022), "English Speech Recognition Method Based on HMM Technology," *2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, P. 646–649. DOI: https://doi.org/10.1109/ICITBS53129.2021.00164
15. Mykhailichenko, I., Ivashchenko, H., Barkovska, O., & Liashenko, O. (2022), "Application of Deep Neural Network for Real-Time Voice Command Recognition", *In 2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek)*, P. 1–4. DOI: https://doi.org/10.1109/KhPIWeek57572.2022.9916473
16. Barkovska, O., Lytvynenko, V., (2022), "Study of the performance of neural network models in semantic analysis", *Modern trends in the development of information and communication technologies and management tools*, Vol.1, P. 136.
17. Barkovska, O., Kholiev,V., Lytvynenko, V. (2022), "Study of noise reduction methods in the sound sequence when solving the speech-to-text problem", *Advanced Information Systems*, No. 6.1, P. 48–54. DOI: https://doi.org/10.20998/2522-9052.2022.1.08
18. Xu, Y., Du, J., Dai, L. -R., and Lee, C. -H. (2015), "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 1, P. 7–19. DOI: https://doi.org/10.1109/TASLP.2014.2364452
19. Davydov, V., & Hrebeniuk, D. (2020), "Development of the methods for resource reallocation in cloud computing systems", *Innovative Technologies and Scientific Solutions for Industries*, 3 (13), P. 25–33. DOI: https://doi.org/10.30837/ITSSI.2020.13.025

Received 03.12.2022

Відомості про авторів / Сведения об авторах / About the Authors

Барковська Олеся Юріївна – кандидат технічних наук, доцент, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна; e-mail: olesia.barkovska@nure.ua; ORCID ID: https://orcid.org/0000-0001-7496-4353

Барковская Олеся Юрьевна – кандидат технических наук, доцент, доцент кафедры электронных вычислительных машин, Харьковский национальный университет радиоэлектроники, Харьков, Украина.

Barkovska Olesia – Ph.D (Engineering Sciences), Docent, Associate Professor Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 4 (22)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

ДОСЛІДЖЕННЯ РОБОТИ МОДУЛЯ ПЕРЕТВОРЕННЯ МОВИ В ТЕКСТ У ЗАПРОПОНОВАНІЙ МОДЕЛІ АВТОМАТИЧНОГО АНОТУВАННЯ ПРОМОВИ СПІКЕРА

Предметом дослідження є модуль конвертації мови спікера в текст у запропонованій моделі автоматичного анотування промови спікера, що стає все більш затребуваним в Україні завдяки активному переходу спілкування, навчання, проходження тренінгів, співбесід, обговорення важливих питань тощо на форму онлайн. **Мета роботи** – скорочення появи хибних або спотворених даних під час перетворення звукового ряду в релевантний текст для подальшого семантичного аналізу. Для досягнення поставленої мети були виконані такі **завдання:** запропоновано узагальнену модель текстової сумаризації вхідних аудіоданих; проналізовано наявні моделі STT (перетворення аудіоданих у текст); досліджено можливість роботи модуля STT з українською мовою; оцінено якість роботи модуля STT та таймінгу роботи з українською та англійською мовами. Запропонована модель автоматичного анотування промови спікера має два головних функціональних модулі: *speech-to-text* (STT) і *summarization module* (SUM). Для модуля STT досліджено та вдосконалено такі **моделі** лінгвістичного аналізу тексту: для англійської мови це *wav2vec2-xls-r-1b3*, а для української – *Ukrainian STT model* (*wav2vec2-xls-r-1b-uk-with-lm*), математичним апаратом яких є нейронні мережі. Отримано такі **результати**: завдяки використанню додаткової української лінгвістичної моделі *wav2vec2* зменшується показник рівня помилок слів майже в 1,5 раза, що впливає на якість розпізнавання слів з аудіо й потенційно може сприяти отриманню більш якісних текстових даних на виході. Для оцінювання таймінгу роботи модуля STT було проаналізовано три аудіозаписи англійською та українською мовами різної довжини: 5 с, ~60 с та ~240 с. Результати показали помітну тенденцію прискорення отримання вихідного файлу за умови використання обчислювального ресурсу графічного прискорювача *NVIDIA Tesla T4* саме для найдовшого аудіозапису. **Висновки**. Використання глибокої нейронної мережі на етапі шумопригнічення у вхідному файлі є виправданим, оскільки забезпечує збільшення метрики WER майже на 25%, а збільшення обчислювальних потужностей графічного процесора та кількості потокових процесорів надають прискорення лише для вхідних аудіофайлів великого розміру. Подальші дослідження автора спрямовані на вивчення ефективності методів модуля сумаризаціїї отриманого тексту.

Ключові слова: STT; текст; оброблення; анотація; реферат; аудіофайл; модель; навчання.

ИССЛЕДОВАНИЕ РАБОТЫ МОДУЛЯ ПРЕОБРАЗОВАНИЯ РЕЧИ В ТЕКСТ В ПРЕДЛОЖЕННОЙ МОДЕЛИ АВТОМАТИЧЕСКОГО АНОТИРОВАНИЯ РЕЧИ СПИКЕРА

Предметом исследования является модуль конвертации речи спикера в текст в предложенной модели автоматического аннотирования речи спикера, который становится все более востребованным в Украине из-за активного перехода общения, обучения, прохождения тренингов, собеседований, обсуждения важных вопросов и т.д. на форму онлайн. **Целью** работы является сокращение появления ложных или искаженных данных при преобразовании звукового ряда в релевантный текст для дальнейшего семантического анализа. Для достижения поставленной цели были решены следующие **задачи**: предложена обобщенная модель текстовой суммаризации входных аудиоданных; пронализированы существующие модели STT (превращение аудиоданных в текст); исследована возможность работы модуля STT на украинском языке; выполнена оценка качества работы модуля STT и тайминга работы на украинском и английском языках. Предлагаемая модель автоматического аннотирования речи спикера имеет два главных функциональных модуля: *speech-to-text* (STT) и *summarization module* (SUM). Для модуля STT исследованы и усовершенствованы следующие **модели** лингвистического анализа текста: для английского языка это *wav2vec2-xls-r-1b3*, а для украинского – *Ukrainian STT model* (*wav2vec2-xls-r-1b-ru-with-lm*), математическим аппаратом которых являются нейронные сети. Получены следующие **результаты**: благодаря использованию дополнительной украинской лингвистической модели *wav2vec2* уменьшается показатель уровня ошибок слов почти в 1,5 раза, что влияет на качество распознавания слов по аудио и потенциально может привести к получению более качественных текстовых данных на выходе. Для оценки тайминга работы модуля STT было проанализировано три аудиозаписи на английском и украинском языках разной длины: 5 с, ~60 с и ~240 с. Результаты показали заметную тенденцию ускорения получения исходного файла при использовании вычислительного ресурса графического ускорителя *NVIDIA Tesla T4* именно для самой длинной аудиозаписи. **Выводы**. Использование глубокой нейронной сети на этапе шумоподавления во входном файле оправдано, поскольку обеспечивает увеличение метрики WER почти на 25%, а увеличение вычислительных мощностей графического процессора и количества потоковых процессоров предоставляют ускорение только для входных аудиофайлов большого размера. Последующие исследования автора сосредоточены на исследовании эффективности методов модификации суммаризации полученного текста.

Ключевые слова: STT; текст; обработка; аннотация; реферат; аудиофайл; модель; обучение.

Бібліографічні описи / Bibliographic descriptions

Барковська О. Ю. Дослідження роботи модуля перетворення мови в текст у запропонованій моделі автоматичного анотування промови спікера. *Сучасний стан наукових досліджень та технологій в промисловості*. 2022. № 4 (22). С. 5–13. DOI: https://doi.org/10.30837/ITSSI.2022.22.005

Barkovska, O. (2022), "Research into speech-to-text tranfromation module in the proposed model of a speaker's automatic speech annotation", *Innovative Technologies and Scientific Solutions for Industries*, No. 4 (22), P. 5–13. DOI: https://doi.org/10.30837/ITSSI.2022.22.005