Speech-To-Text Software Design for the High Education Learning

Elena Yashina, Tetiana Rubanik and Andriy Chukhray

National Aerospace University «Kharkiv Aviation Institute», Chkalova St., 17, 61070 Kharkiv, Ukraine

Abstract

The article is dedicated to the application of Speech-to-text technologies and software in the high education. The computer-aided and learning assistive technologies plays an important role in the educational process of a modern university both in online and face-to-face learning. The article considers the use of Automatic speech recognition and Speech-to-text technology and software in the education. Reviewed available Speech-to-text services. The architecture design for an application for the preparing of lectures notes is proposed. The need to integrate the application with other learning assistive technologies (text editors, graphical editors, Image-to-text and other) is noted. Difficulties of Speech-to-text technologies applying and development perspectives are considered.

Keywords [1]

Speech-to-text, computer-aided learning, learning assistive technologies.

1. Introduction

Modern universities actively use various e-learning and learning assistive technologies [1]. One of them is Automatic speech recognition and Speech-to-text technology that allows you to quickly get a recording of a lecture or transcribe a pre-made audio recording. Initially, these technologies were used in teaching special categories of learners: hearing impaired or foreign students. However, the progress of technology, growth in the speed and accuracy of conversion, as well as the emergence of accessible services have made it possible to significantly expand the scope of their application. Providing information in different forms: audial, visual, textual allows better to meet the needs of students with different learning styles

The COVID-19 pandemic has caused a massive shift to online learning [2]. The role of digital learning platforms has dramatically increased. Various resources, including lecture recordings, should be available to students at any time.

Even more significant changes occurred after the beginning of the war in Ukraine in 2022. Thousands of students lost access to classrooms and libraries of their universities. Synchronized online sessions interrupted by air raid alerts and blackouts. Asynchronous learning mode becomes a priority. This places high requirements on the completeness and availability of educational resources. The combined impact of the pandemic and war has had a destructive impact on all areas of life [3]. Overcoming the encountered problems is impossible without the digitalization of various areas, first of all education.

Usually the lecture is delivered in the form of a video recording. But the presence of text transcript provides additional opportunities. The text is easy to edit, supplement, correct inaccuracies. It is easy to give a text document a structure and a table of contents. It is possible to search by keywords, etc. However, lectures on technical and natural sciences are full of illustrative material: formulas, drawings, diagrams, etc. Turning a text transcript into a full-fledged lecture notes requires the use of additional editing tools.

2nd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2022), December 2-4, 2022, Łódź, Poland
EMAIL:o.yashina@khai.edu (E. Yashina); t.m.rubanik@student.khai.edu (T. Rubanik); achukhray@gmail.com (A. Chukhray)
ORCID: 0000-0003-2459-1151 (E. Yashina); 0000-0002-8277-1575 (T. Rubanik); 0000-0002-8075-3664 (A. Chukhray)
© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3348/paper4.pdf

The aim of this article is improving the availability of educational resources such as lecture records via the Speech-to-text software development and implementation.

2. Automatic speech recognition and Speech-to-text technology in the education

Speech-to-text (STT) technology was previously mainly used to help certain groups of students (i.e., students with learning or physical disabilities or foreign students) in order to guarantee them the equal access to learning. However, as time passed by, the target users involved into research on STT technology has got broader. That is, nowadays STT technology is adopted to assist not only students with special needs but also general population of students for more educational purposes, such as enhancing students' understanding of a presented learning content during and after academic activities

The paper [4] reviewed literature from 1999 to 2014 inclusively on how Speech-to-Text Recognition technology has been applied to enhance learning. Four main areas of use of STT in teaching have been identified.

1. Students with cognitive or physical disabilities. It is extremely difficult for these students to focus their visual attention on note-taking and the instructor (or interpreter) simultaneously. Therefore, it was suggested to apply assistive technologies, such as a speech-to-text support service, to enhance computer-assisted learning for students with different types of disabilities.
2. Non-native speakers and foreign students. The speech-to-text recognition technology is a potentially reliable tool for non-native speaker students to better understand a speech given in a foreign language.
3. Online students. Network traffic congestion can cause poor quality of audio communication in a synchronous cyber classroom. Under such condition, students are not able to hear a speaker clearly. This issue was viewed as one technological challenge. It negatively affects online teaching and learning activities as it hinders students' understanding of a delivered speech, and it also hampers students from engaging in classroom participation and interaction. STT technology allows you to create real-time transcripts or subtitles that are simultaneously displayed to students on their computer screens. In this way Students could listen to the speaker and read transcripts at the same time. More importantly, the text generated by STT was saved for further revision to fix some recognition errors, and students could get a near verbatim transcript to review after class.
4. Students in traditional learning environment. the adoption of the STT technology in traditional learning environment has several benefits. One of them is to improve teaching methods and to enhance learning opportunities. For example, by using the STT, teachers can take a proactive, rather than a reactive approach to teach students with different learning styles. It provides educators with a practical means of making their teaching accessible and improves the quality of instruction in the process.

Two different methods of STT-mediated lecture absorption are most commonly used, such as real-time subtitles and transcription after the lecture When lecture transcripts were available, students were able to pay more attention to the instructor instead of focusing on recording complete class notes, and with the lecture transcripts, they could review the lecture material for several times. Besides, students were able to take notes, make comments and remarks, and look for specific text by searching keywords and time periods. The students who had access to post lecture transcriptions received higher scores. However, most students claimed that the accuracy rate of STT technology was not precise enough, and text generated with many errors could distract their attention from the lecture.

Review [4] shows that participants in most studies on STT, no matter what category of users they belong to and no matter what learning environment they learn in, had positive perceptions toward usefulness of STT transcripts for learning.

According to survey [5] 20% of British higher education students using various assistive technologies in 2020. 9% of students using dictation (speech to text) use technology. 51% students says their organization had offered them support in using assistive technologies. Approximately half of students (54%) said they enjoyed trying out new and innovative technologies, and less than half (43%) of students said they were comfortable using mainstream technologies. 89% of higher education students had access to online course materials, e-books and journals and recorded lectures at their organization whenever they needed them.

The 2020/21 survey [6] was taken at a time when students, faculty and colleges continued to experience disruptions caused by the COVID19 pandemic. Colleges were supposed to respond quickly to a changing environment to maintain and reimagine training and support they were able to suggest and also solve many other operational aspects of delivery. 58% of students evaluated quality of online learning materials as well designed although less than half of learners (41%) agreed that the online learning materials were engaging and motivating. Two thirds of learners (66%) had accessed course materials and notes. Substantial numbers (63%) had also submitted coursework, taken part in live online lectures/teaching sessions.

Learners were asked to say what they thought were the most positive and negative aspects of online learning. Their responses reveal that learning preferences are very individual – what some learners really like, others do not. Lecture recordings which is interesting as lecture recordings have not traditionally been so available/used. Students also noted easy and convenient access to learning resources, materials and information.

The study [7] shows that students' attendance and engagements have significantly dropped during live online delivery due to the impact of the COVID-19 pandemic. However, during the pandemic the way technology has been used to deliver learning using recorded lectures and seminars on a virtual platform, attendance and engagement in higher education seem to lose their importance since students do not have to attend classes to get access to the course material.

Recording lectures and seminars became the norm in higher education for synchronous and asynchronous teaching during COVID-19. The students who was graduated in 2020-21 are much better prepared than last year's graduating students. These new cohorts of students have a very clear understanding of the role and implications of the online recorded lectures and seminars on their learning and knowledge.

3. Speech-to-text methods and technologies

When sounds come out of someone's mouth to create words, it creates a series of vibrations. Speech-to-text technology works by picking up these vibrations and translating them into digital language using an analog-to-digital converter. An analog-to-digital converter takes sounds from an audio file, measures the waves in great detail, and filters them to distinguish the corresponding sounds. Sounds are broken down into hundredths or thousandths of a second and then matched with phonemes. A phoneme is a sound unit that distinguishes one word from another in any given language. For example, there are approximately 40 phonemes in the English language. Once broken down, the phonemes are run through a mathematical model that compares them to known sentences, words and phrases. The result is provided as text based on the most likely version of the audio.

A computer program uses linguistic algorithms, such as Automatic speech recognition (ASR), to sort audio signals from spoken words and translate those signals into text. Speech-to-text works using a complex machine learning model that involves several steps [8].

3.1. Automatic speech recognition methods

The function of an ASR is to take input of a sound wave and convert the spoken speech into text form; the input could be either taken directly using a microphone or as an audio file [9]. This multimedia tools and applications problem can be explained in the following way: for a given sequence input sequence X, where $X = X_1, X_2, \ldots., X_n$, where n is the length of the input sequence, the function of an ASR is to find a corresponding output sequence Y, where $Y = Y_1, Y_2, \ldots., Y_m$, where m is the length of the output sequence. And the output sequence Y has the highest posterior probability $P(Y|X)$, where $P(Y|X)$ can be calculated using the given formula:

$$W = \text{argmax} \frac{P(W)P(X|W)}{P(X)}, \qquad (1)$$

where $P(W)$ is the probability of the occurrence of the word, $P(X)$ is the probability that X is present in the signal, and $P(X|W)$ is the probability of the acoustic signal W occurring in correspondence to the word X.

An ASR can generally be divided into four modules: a pre-processing module, a feature extraction module, a classification model, and a language model. Usually the input given to an ASR is captured using a microphone. This implies that noise may also be carried alongside the audio. The goal of preprocessing the audio is to reduce the signal-to-noise ratio. There are different filters and methods that can be applied to a sound signal to reduce the associated noise. Framing, normalization, end-point detection and pre-emphasis are some of the frequently used methods to reduce noise in a signal. Preprocessing methods also vary based on the algorithm being used for feature extraction. Certain feature extraction algorithms require a specific type of pre-processing method to be applied to its input signal. After pre-processing, the clean speech signal is then passed through the feature extraction module. The performance and efficiency of the classification module are highly dependent upon the extracted features.

There are different methods of extracting features from speech signals. Features are usually the predefined number of coefficients or values that are obtained by applying various methods on the input speech signal. The feature extraction module should be robust to different factors, such as noise and echo effect. Most commonly used feature extraction methods are Melfrequency cepstral coefficients, linear predictive coding, and discrete wavelet transform.

The third and final module is the classification model; this model is used to predict the text corresponding to the input speech signal. The classification models take input of the features extracted from the previous stage to predict the text. Like the feature extraction module, there are different types of approaches that can be applied to perform the task of speech recognition.

The first type of approach uses joint probability distribution formed using the training dataset, and that joint probability distribution is used to predict the future output. This approach is called a generative approach; hidden Markov model and Gaussian mixture models are the most commonly used models based on this approach.

The second approach calculates a parametric model using a training set of input vectors and their corresponding output vectors. This approach is called the discriminative approach; support vector machines and artificial neural network are its most common examples. Hybrid approaches can also be used for classification purposes; one example of such a hybrid model is that of a hidden Markov model and artificial neural network. The language model is the last module of the ASR; it consists of various types of rules and semantics of a language. Language models are necessary for recognizing the phoneme predicted by the classifier; and is also used to form trigrams, words or sentences using all of the predicted phonemes of a given input. Most modern ASRs are designed to work without Language Models as well. Such ASRs can predict words and sentences spoken in the given input, but their efficiency can be increased significantly by using a language model.

3.2. ASR accuracy evaluation

Evaluation is one of the most important aspects of a conducted research because of its importance this section explains in detail different metrics that can be used to evaluate the performance of an ASR. The performance of a speech recognition system usually depends on two factors, the accuracy of the output produced as well as the processing speed of the ASR.

The following methods can be used to measure the accuracy of an ASR. The accuracy of an ASR is hard to calculate as the output produced by the ASR may not have the same length as the ground truth. Word error rate (WER) is the commonly used metric to estimate the performance of an ASR, as it calculates error on word level rather than phoneme level [10]. The WER can be calculated using the following formula:

$$WER = \frac{S + D + I}{N}, \tag{2}$$

where *S* is the number of substitutions performed in the output text as compared to the ground truth, *D* is the number of deletions performed, and *I* is the number of insertions performed. *N* is the total number of words in the ground truth. Word recognition rate Word Recognition Rate (WRR) is a variation of WER that can also be used to evaluate the performance of an ASR. It can be calculated using the following formula:

$$WRR = 1 - WER. \tag{3}$$

Other metrics is Match error rate (MER)

$$MER = \frac{S + D + I}{H + S + D + I}, \tag{4}$$

$$H = N - (S + D), \tag{5}$$

where parameters *H*, *S*, *D* and *I* correspond to the total number of word hits, substitutions, deletions and insertions [11].

Word information lost (WIL) and Word information preserve (WIP) calculating by formulas:

$$WIP = \frac{H}{N_1}\frac{H}{N_2}, \tag{6}$$

$$WIL = 1 - \ WIP. \tag{7}$$

where N_1 and N_2 are respectively the number of words in groundtruth text and the output transcripts.

The lower are WER, MER and WIL, the better the performance is [12].

Different from WER, BLEU can evaluate whether the transcription maintains the context and organization of the sentence. BLEU was originally proposed for neural machine translation and it claims to be highly correlated with human assessment. BLEU is based on the precision of n-grams, which compares the n-grams of reference text T^* with the n-grams of its transcription *T*. Let be $NG(n, T)$the set of n-grams of text t, the n-gram precision Pn (Equation 2) between texts T^* and *T*:

$$P_n = \frac{|NG(n, T^*) \cap NG(n, T)|}{NG(n, T)}. \tag{8}$$

BLEU is calculated as the geometric mean of P_n, for *n* = 1, 2, 3, 4 multiplied by a factor that penalizes transcriptions shorter than the referenced text. The $BLEU_{penalty}$ factor is 1 if |*T*| > |T^*| and $e^{1-|T^*|/|T|}$, otherwise. BLEU is defined by formula:

$$BLEU = \sqrt[4]{P_1 P_2 P_3 P_4}. \times BLEU_{penalty} \tag{9}$$

METEOR was proposed to fix limitations of BLEU, such as the fact that it does not require explicit word-to-word matching. Another limitation is that its score results in zero whenever one of the n-gram precision is zero, which means the score at sentence level can be meaningless. METEOR is based on the harmonic mean of unigram precision and recall, multiplied by a penalty factor. METEOR defined by formulas:

$$R_n = \frac{|NG(n, T^*) \cap NG(n, T)|}{NG(n, T^*)}, \tag{10}$$

$$METEOR = \frac{10P_1R_1}{R_1 + 9P_1} METEOR_{penalty}. \tag{11}$$

To calculate the $METEOR_{penalty}$, the unigrams in $NG(n, T^*) \cap NG(n, T)$are grouped in chunks, such as each chunk has the maximum number of unigrams in adjacent positions in both T^* and T. The fewer the chunks, the better system transcription matches with the reference transcription [13]:

$$METEOR_{penalty} = \frac{1}{2} \frac{\#chunks}{|NG(n, T^*) \cap NG(n, T)|}. \quad (12)$$

These metrics allow you to evaluate various aspects of the ASR accuracy and performance.

3.3. Speech-to-text tools

The Speech-to-text software recognize and translate spoken language into text using ASR and other computational linguistics. This technology is directly related to computer speech recognition (voice recognition). Certain applications, tools, and devices can transcribe audio streams in real-time to display and interact with text using Unicode characters. Existing speech recognition platforms provide APIs to developers, executing received requests on their own servers, allowing them to be used as "black boxes".

The Speech-to-Text, interprets the words from the user as audio, and converts them to text in written form by utilizing deep learning techniques. Lots applications with automated speech recognition or Speech-To-Text over the past few years have been developed. Thanks to the substantial development of deep neural network, the performance of STT has been drastically improved.

The widely used commercial ASR online services are: Google Cloud Speech-to-Text, is integrated in the widely used platform Google Cloud; IBM Watson Speech-to-Text; Microsoft Azure Cognitive Speech Services; Amazon Transcribe and other.

In the study [12] the accuracy and efficiency of the widely used commercial ASR online services was investigated by metrics WER by formula (2), MER by formula (4), WIP by formula (6) and in the study [13] its accuracy and efficiency was evaluated by metrics WER by formula (2), BLUE by formula (9) and METEOR by formula (11). The main method of evaluating ASR engines is finding out their Word Error Rate. Generally speaking, the lower the WER, the more accurate the speech recognition is. According to this studies, the accuracy of the models by WER is:

1. Google Cloud Speech-to-Text: 11.58– 20.00%
2. IBM Watson Speech-to-Text: 14.81% – 28.57%
3. Amazon Transcribe: 10.27 – 20.00%
4. Microsoft Azure Cognitive Speech Services: 8.14 – 11.11%

Accuracy depends on the dataset used, languages and the presence of noise . Models have comparable accuracy. The choice of service is based on the availability of the API, its functionality, the possibility of integration with other services, etc. Amazon Transcribe service was chosen to develop the system

Amazon Transcribe is part of Cloud Computing Services from Amazon Web Services (AWS). Amazon Transcribe uses a deep learning model to perform ASR to quickly and accurately convert speech to text. In this conversion, the data needs to be first uploaded to Amazon Simple Storage Service (Amazon S3). Then Transcribe calls the objects from S3 for transcription. Though Transcribe jobs can be treated on batch mode (up to 100 parallel jobs). The model provided automatically adds punctuation and number formatting, so that the output closely matches the quality of manual transcription at a fraction of the time and expense. Numbers are also transcribed into digits or “normal form” instead of words.

4. Proposed application design

4.1. Application architecture and tools

The application will have two main functions: speech to text translation and file editing. In order to improve the understanding of the material after conversion into text, it must be edited::

- make corrections to sentences that the system recognized incorrectly;

- break down the material by structure - separate sections and subtopics, etc.;
- add formatting;
- select and add graphic materials to illustrate difficult points, etc.

Software with editor functions of the corresponding file type is suitable for performing the specified steps. Some platforms provide their own Application Program Interfaces (API), Software Developer Kits (SDK), and other ways to integrate ready-made solutions into a new project. The SDK provides a comprehensive collection of tools to create a flexible, intuitive interface.

Amazon Transcribe (aws.amazon.com/transcribe/) — is an automatic speech recognition service that makes it easy to add speech-to-text capabilities to any application. The platform supports speech-to-text conversion both in real time and from pre-recorded audio files, recognition of interlocutors, determination of the language in which the interview is conducted, etc. It has API, communication with S3 storage, SDK for multi-language programming.

In order to interact with the Amazon Transcribe API, it is convenient to use the appropriate client classes from the SDK package.

Language to text translation takes time. Due to this, a call to the API in synchronous mode will stop the entire application for at least half a minute, which makes such an approach inappropriate. The "queue" mechanism was chosen as a means of asynchronous work. Thus, the application will send the request asynchronously, notify the user that their audio has been accepted for processing, and return the control flow.

Figure 1 shows the algorithm for working with Amazon Transcribe using classes from the SDK.

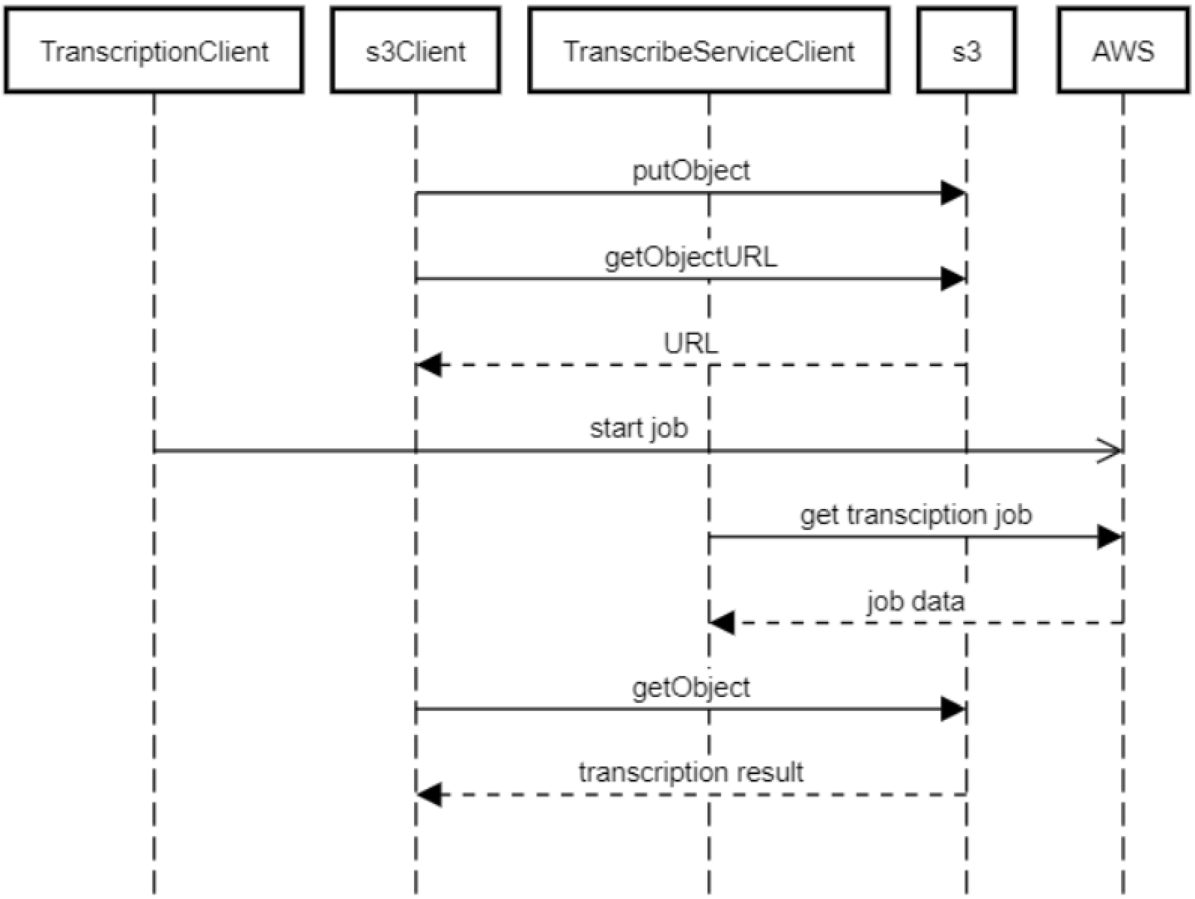


Figure 1: The algorithm for interacting with AWS to extract text from audio

User data is forwarded to the server, processed and stored in a MySQL database. The system uses asynchronous requests that will check the status of `TranscriptionJob` with a time interval of 10-20 seconds. Thus, the system will not waste resources on maintaining the connection and will be able to provide the user with the result with minimal delay. Figure 2 shows the interaction of the developed system with the Amazon service.

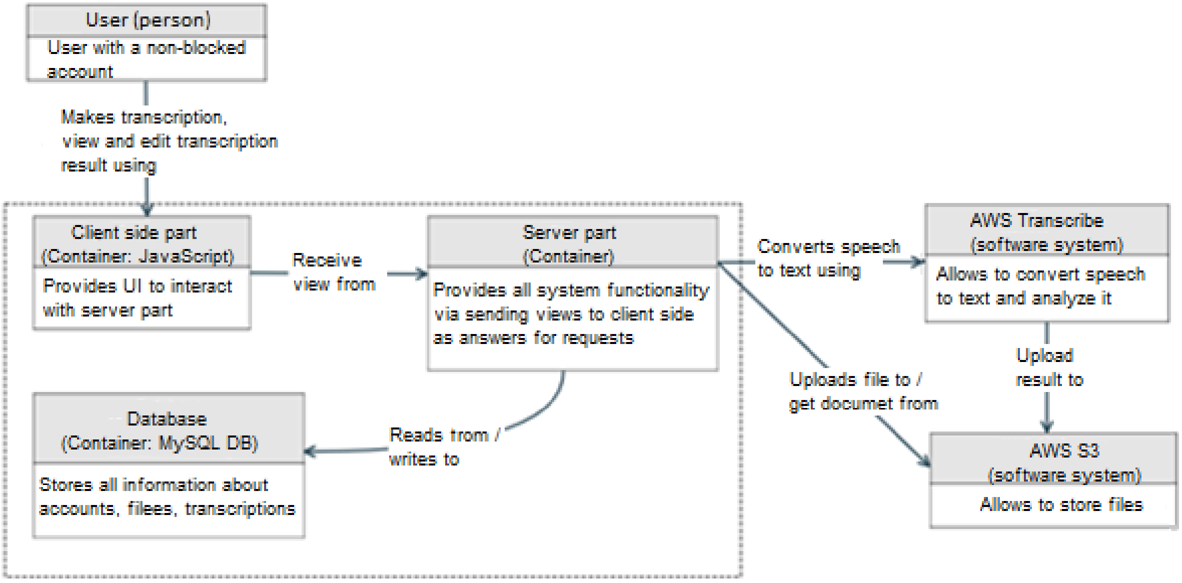

User (person)
User with a non-blocked account
Makes transcription, view and edit transcription result using
Client side part (Container: JavaScript)
Provides UI to interact with server part
Receive view from
Server part (Container)
Provides all system functionality via sending views to client side as answers for requests
Converts speech to text using
AWS Transcribe (software system)
Allows to convert speech to text and analyze it
Upload result to
Uploads file to / get documet from
AWS S3 (software system)
Allows to store files
Reads from / writes to
Database (Container: MySQL DB)
Stores all information about accounts, filees, transcriptions

Figure 2: Interaction of the application under development with Amazon services

The tools to be used during the further development of the application were selected, in particular:

- Synfony as a PHP framework (symfony.com);
- Vue as a JavaScript framework (vuejs.org);
- Docker as an environment modeling tool: application server, database server, etc. (www.docker.com);
- MySql as a database (www.mysql.com);
- Redis as a database for queue management (redis.com);
- Supervisord as a client/server system that allows you to monitor a number of processes in UNIX-like operating systems (supervisord.org);
- AWS SDK PHP as a set of tools for interacting with Amazon (github.com/aws/aws-sdk-php);

4.2. Application structure

Symfony was chosen as the framework for application development. This dictates the choice of a pattern for interacting with the database: a layer of repository classes that emulate `ServiceEntityRepository`. Entity classes will accordingly be collected in a separate group - `Entity`.

Services will be the layer containing business logic (Figure 3). This will separate the application logic from the models and controllers (thin controller pattern). Algorithms to be executed asynchronously will request handler classes that implement the `MessageHandlerInterface` provided by Symfony. `Messenger` centers around two different classes: the message class that contains the data, and the handler class that will be called when that message is sent. The handler class will read the message class and perform one or more tasks. The system under development will have two handlers: for the request to start transcription and for the request for the status of the started transcription (Figure 4).

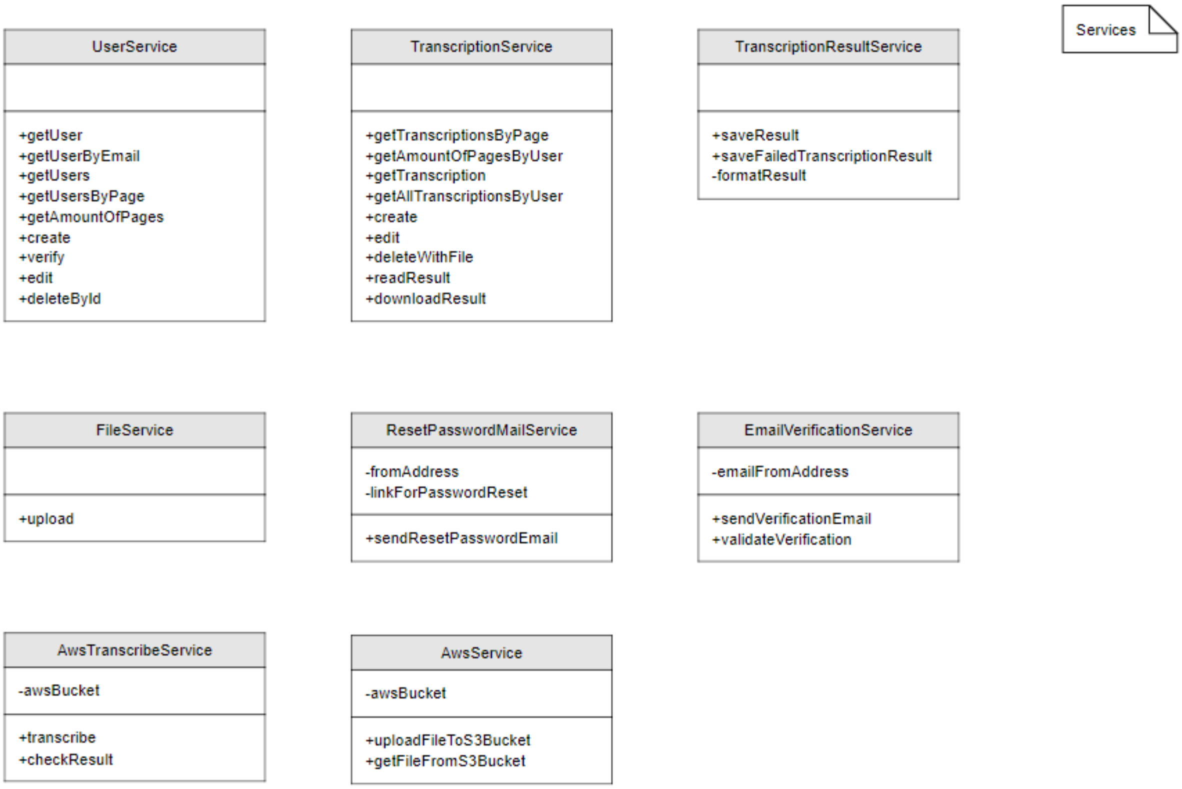


Figure 3: System services for the implementation of the functionality of the speech-to-text application

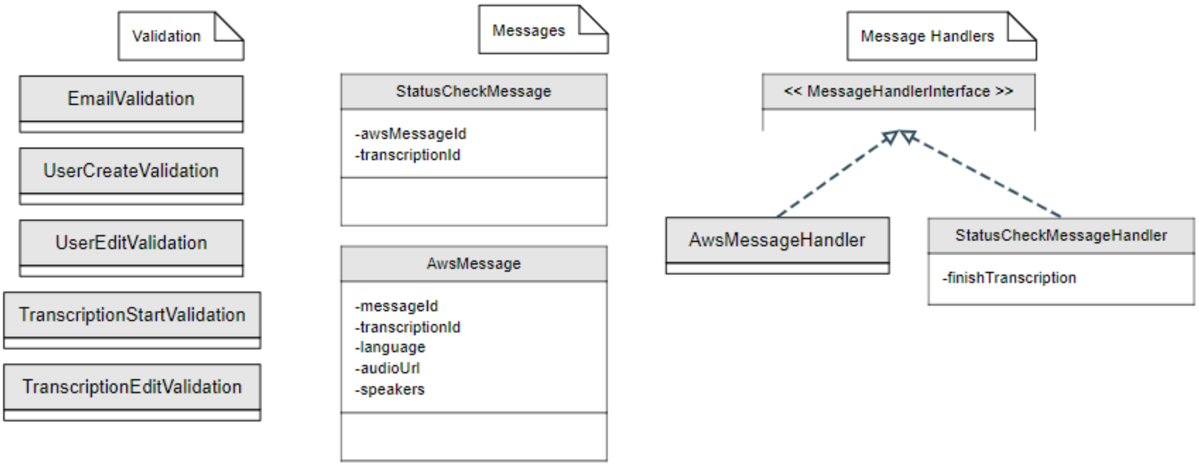


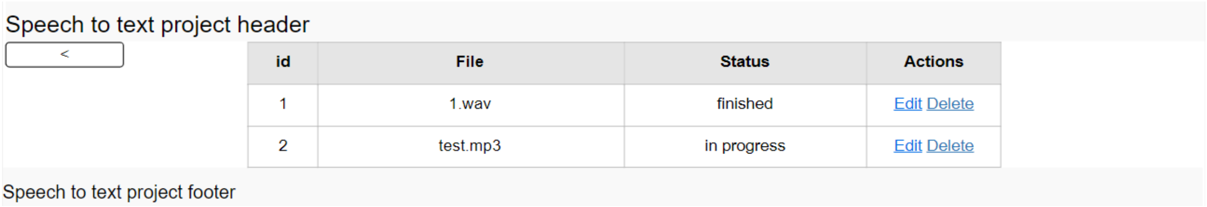
Figure 4: Handler classes, corresponding storage classes, and validator classes

Management of access to the system, control of access to functions according to the user's role will be carried out by classes from the `Security` group. Symfony is delivered with many authenticators, and third-party packages also implement more complex cases such as JWT and oAuth 2.0. The application will use `JWTAuthenticator` and a separate authenticator for the first login and receiving the access token. The `Voter` component provided by the framework allows you to implement access restrictions. According to the entities and roles of the account, the user will be blocked (return an error with a reference to the access level) from some functions of the controllers (viewing accounts, downloading the transcription result, etc.).

4.3. User interface design

The layout of the application interface has been developed.

After logging into the account, the user should see the transcription creation page. The next step will be viewing the list of audio files with existing transcriptions (Figure 5).



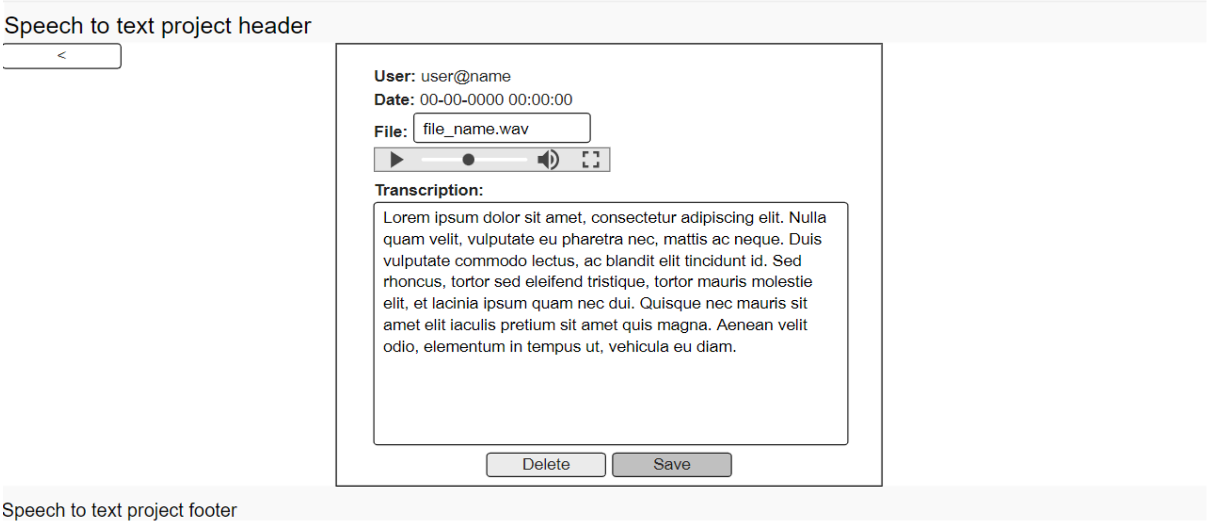
Speech to text project header

id	File	Status	Actions
1	1.wav	finished	Edit Delete
2	test.mp3	in progress	Edit Delete

Speech to text project footer

Figure 5: A list of transcriptions

The user interface for the lecture file editing section will be built using the SDK. Pdftron editor (www.pdftron.com) and related products are choiced. The editing window will correspond to the demo displayed on the official page of the tool (Figure 6).



Speech to text project header

User: user@name

Date: 00-00-0000 00:00:00

File: file_name.wav

Transcription:

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus, ac blandit elit tincidunt id. Sed rhoncus, tortor sed eleifend tristique, tortor mauris molestie elit, et lacinia ipsum quam nec dui. Quisque nec mauris sit amet elit iaculis pretium sit amet quis magna. Aenean velit odio, elementum in tempus ut, vehicula eu diam.

Speech to text project footer

Figure 6: Transcription editing

The user can edit the transcription text by adding headings, links and graphic elements (Figure 7).

The prepared text can be converted to a convenient format (for example, PDF) and placed on the learning management system (Figure 8).

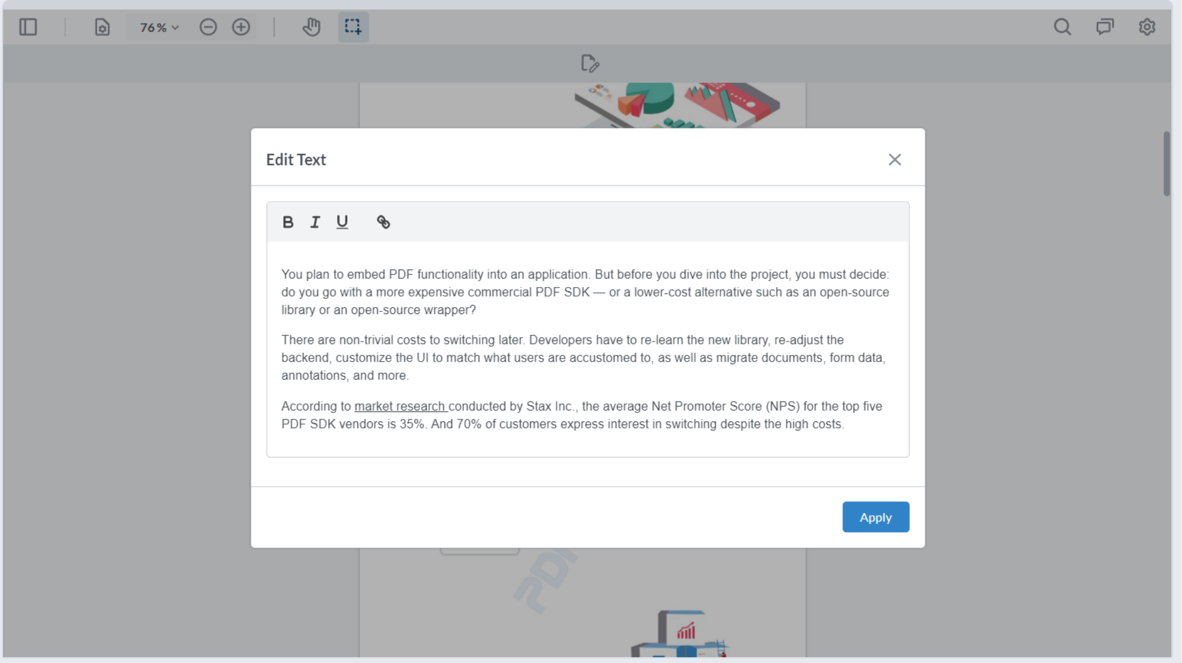
5. Conclusions and perspectives

Educational assistive technology significantly improve the quality and effectiveness of learning both online and face-to-face. SST allows to speed up the preparation and delivery of educational resources to learning management system. The lectures transcription allows students to better understand the topic.

However, you need to keep in mind the disadvantages of Speech-to-text:

- insufficient performance of existing systems;
- the transcript contains some errors and inaccuracies;
- the quality of transcription depends on the clarity of pronunciation, requires a good microphone and the absence of external noise;

- the best transcription quality is provided for English, the recognition of Ukrainian is much worse;
- transcription of a discussion with several participants can be worse than a lecture by a single teacher.

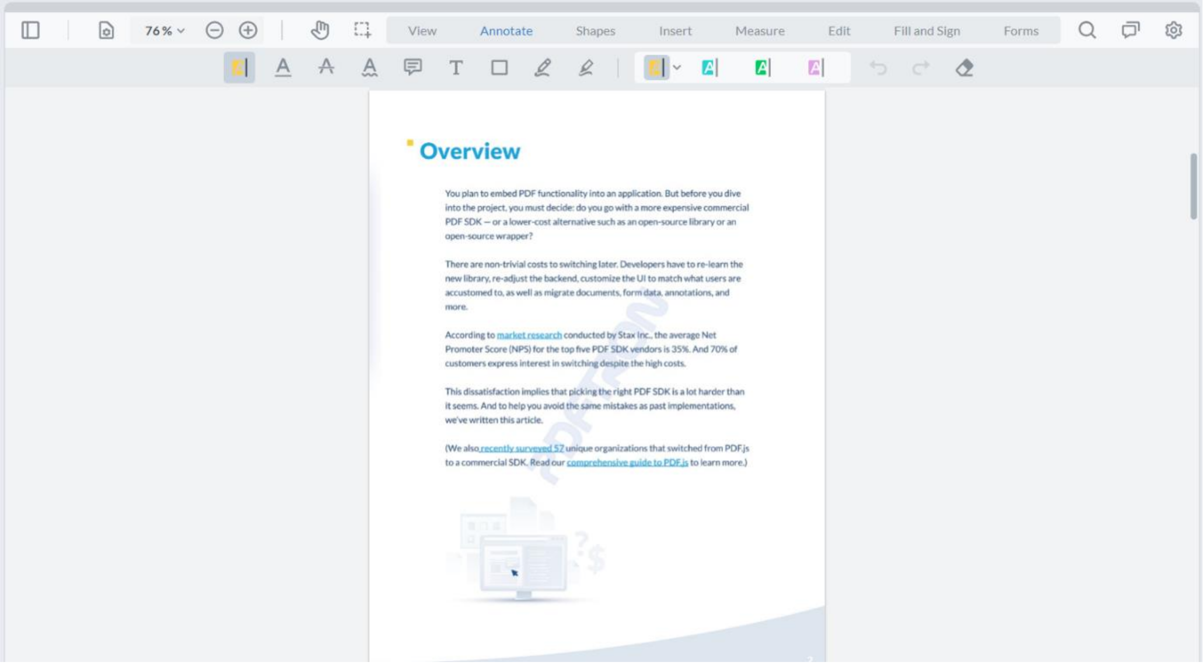


You plan to embed PDF functionality into an application. But before you dive into the project, you must decide: do you go with a more expensive commercial PDF SDK — or a lower-cost alternative such as an open-source library or an open-source wrapper?

There are non-trivial costs to switching later. Developers have to re-learn the new library, re-adjust the backend, customize the UI to match what users are accustomed to, as well as migrate documents, form data, annotations, and more.

According to market research conducted by Stax Inc., the average Net Promoter Score (NPS) for the top five PDF SDK vendors is 35%. And 70% of customers express interest in switching despite the high costs.

Figure 7: Editing the text page



Overview

You plan to embed PDF functionality into an application. But before you dive into the project, you must decide: do you go with a more expensive commercial PDF SDK — or a lower-cost alternative such as an open-source library or an open-source wrapper?

There are non-trivial costs to switching later. Developers have to re-learn the new library, re-adjust the backend, customize the UI to match what users are accustomed to, as well as migrate documents, form data, annotations, and more.

According to market research conducted by Stax Inc., the average Net Promoter Score (NPS) for the top five PDF SDK vendors is 35%. And 70% of customers express interest in switching despite the high costs.

This dissatisfaction implies that picking the right PDF SDK is a lot harder than it seems. And to help you avoid the same mistakes as past implementations, we've written this article.

(We also recently surveyed 57 unique organizations that switched from PDF.js to a commercial SDK. Read our comprehensive guide to PDF.js to learn more.)

Figure 8: Document prepared for publication

Promising areas of research and development are related to the improvement of technologies, as well as the development of means for integrating various tools. The proposed system will be used in conjunction with other available software tutoring tools [14, 15]

Speech-to-Text, Text-to-Speech, Image-to-Text integration will significantly improve efficiency of e-learning technology [16]. In turn, the integration of assistive technologies based on a unified digital platform will significantly expand the possibilities of learning management system.

However, it must be taken into account that training in an e-learning environment happens differently than in the traditional classroom and can present new challenges to teachers and students in this online learning environment [17]. The introduction of tools requires not only the availability of hardware and software, but also significant changes in the culture and style of teaching, enhancing the digital experience of teachers and students.

6. References

[1] D.A. Huffaker, S.L. Calvert., The new science of learning: Active learning, metacognition, and transfer of knowledge in e-learning applications, Journal of Educational Computing Research 29, no. 3 (2003): 325-334. doi: 10.2190/4T89-30W2-DHTM-RTQ2

[2] V.J. García-Morales, A. Garrido-Moreno, R. Martín-Rojas, The transformation of higher education after the COVID disruption: Emerging challenges in an online learning scenario, Frontiers in Psychology 12 (2021): 616059. doi: 10.3389/fpsyg.2021.616059

[3] D. Chumachenko, P. Pyrohov, I. Meniailov, T. Chumachenko, Impact of war on COVID-19 pandemic in Ukraine: the simulation study, Radioelectronic and Computer Systems 2 (2022): 6-23. doi: 10.32620/reks.2022.2.01

[4] R. Shadiev, W.Y. Hwang, N.S. Chen, Y.M. Huang, Review of speech-to-text recognition technology for enhancing learning, Journal of Educational Technology & Society 17, no. 4 (2014): 65-84. URL: https://www.jstor.org/stable/jeductechsoci.17.4.65

[5] M. Langer-Crame, C. Killen, H. Beetham, Student digital experience insights survey 2020: question by question analysis of findings from students in UK further and higher education, (2020). URL: https://www.voced.edu.au/content/ngv:91499

[6] Student digital experience insights survey 2020/21 [findings from pulse 1: October-December 2020], Bristol, England: JISC, 2021. URL: https://www.voced.edu.au/content/ngv:91498

[7] S. Ghosh, Y. Liang, Recorded teaching materials and their impact on students attendance, engagement and performance during Covid-19, Academy of Marketing Conference 2021: Reframing Marketing Priorities, 05-07 July 2021, Virtual (2021). URL: https://eprints.bournemouth.ac.uk/35731/

[8] M. Anniss, How Does Voice Recognition Work?. The Rosen Publishing Group, 2013.

[9] M. Malik, M.K. Malik, K. Mehmood, I. Makhdoom, Automatic speech recognition: a survey, Multimedia Tools and Applications 80, no. 6 (2021): 9411-9457. doi: 10.1007/s11042-020-10073-7

[10] B. Favre, K. Cheung, S. Kazemian, A. Lee, Y. Liu, C. Munteanu, A. Nenkova et al, Automatic human utility evaluation of ASR systems: Does WER really predict performance?, INTERSPEECH, pp. 3463-3467. 2013. URL: https://pageperso.lis-lab.fr/~benoit.favre/papers/favre_interspeech2013.pdf

[11] A.C. Morris, V. Maier, P. Green, From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition, Eighth International Conference on Spoken Language Processing 2004. URL: https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2004/i04_2765.pdf

[12] B. Xu, C. Tao, Z. Feng, Y. Raqui, S. Ranwez, A benchmarking on cloud based speech-to-text services for french speech and background noise effect, APIA 2021 - Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (événement affilié à PFIA 2021), Jun 2021, Bordeaux, France. p. 102-107. URL: https://hal.mines-ales.fr/hal-03277773

[13] R.P. Magalhães, D.J.R. Vasconcelos, G.S. Fernandes, L.A. Cruz, M.X. Sampaio, J.A.F. de Macêdo, T.L.C. da Silva, Evaluation of Automatic Speech Recognition Approaches, Journal of Information and Data Management 13, no. 3 (2022). doi: 10.5753/jidm.2022.2514

[14] A. Chukhray, O. Havrylenko, The engineering skills training process modeling using dynamic bayesian nets, Radioelectronic and Computer Systems 2 (2021): 87-96. doi: 10.32620/reks.2021.2.08.

[15] A. Chukhray, E. Yashina, Models and Software for Intelligent Web-Based Testing System in Mathematics, CEUR Workshop Proceedings 3003 (2021): 1-10. 2021. URL: https://ceur-ws.org/Vol-3003/paper1.pdf

[16] A.S. Deshpande, S.V. Shreyas, P.B. Swami, P.R. Jaiswal, Integration of Speech, Image & Text Processing Technologies (2017): 251-254. URL: https://www.academia.edu/download/53485551/IRJET-V4I450.pdf

[17] A.M. Tirziu, Andreea-Maria, C. Vrabie, Education 2.0: E-learning methods, Procedia-Social and Behavioral Sciences 186 (2015): 376-380. doi: 10.1016/j.sbspro.2015.04.213