UDC 621.3

DOI: 10.20535/2523-4455.mea.255961

Development of a System and Interface for Speech Synthesis in Ukrainian for Websites

O. R. Osadchuk, ORCID 0000-0003-4934-2565
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute» ROR 00syn5v21
Kyiv, Ukraine

***Abstract*—The paper describes the system of content synthesis and sounding on websites in Ukrainian, designed to simplify the consumption of content for the visually impaired, which features easy integration into the most popular content management system on sites, namely CMS WordPress.**

Currently, people with visual impairments are severely limited in their use of Internet products because most web resources are not tailored to their needs. Modern information technology allows such people to receive information along with healthy ones thanks to solutions developed by scientists and engineers from different countries.

Text information can be delivered to the visually impaired with a magnifying glass, or by enlarging the font by software, blind - by sounding the text using computer programs or displaying texts on the Braille screen of the monitor. This is an effective solution, but the choice of methods for reproducing such information must be provided fully by people with disabilities themselves, which is a significant problem due to the significant time spent on information consumption.

To facilitate the perception of visually impaired people when using websites, an international standard for webmasters has been developed - a guide to web content accessibility called the W3 Web Content Accessibility Guidelines (WCAG) Consortium 2.0. The standard describes in detail the requirements for visually impaired people that are recommended to be met in order for them to view the website without any problems. The basic provisions of WCAG 2.0 will describe the parameters and algorithms for scaling, clustering and separation of information by programs for the visually impaired and provide recommendations for writing website code.

However, to implement such recommendations, webmasters need to learn new programming principles and algorithms and use additional development tools. This is often difficult, requiring additional training, which entails non-compliance by webmasters with such requirements.

Keywords — speech synthesis; support low vision; recognition algorithm; natural language processing; neural network; understanding of natural language; web integration; CMS Wordpress.

1. *Introduction*

Modern information technology allows visually impaired people to receive information along with healthy people through a number of technical solutions, but the choice of methods of reproducing such information must be fully provided by people with disabilities, this is a significant problem due to significant time spent on information. Modern information technology allows such people to receive information along with healthy people thanks to solutions developed by scientists and engineers from different countries [1].

To facilitate the perception of visually impaired people when using websites, an international standard for webmasters has been developed, a guide to web content accessibility called the 2.0 consortium W3 Web Content Accessibility Guidelines (WCAG) [2]. The standard describes in detail the requirements for visually impaired people that are recommended to be met in order for them to view the website without any problems. The main provisions of WCAG 2.0 will describe the parameters and algorithms for scaling, clustering and separation of information by programs for the visually impaired and provide recommendations for writing website code [3].

However, to implement such recommendations, webmasters need to learn new programming principles and algorithms and use additional development tools. This is often difficult, requiring additional training, which entails non-compliance by webmasters with such requirements. The article describes the system of synthesis of Ukrainian speech, easy for webmasters, in the installation, configuration, and use for consumption of content on web pages for the visually impaired.

The idea of the project lies in the development of a simple and easy package solution for ensuring the possibility of voiceover text on a website for people from all over the world. The decision is guilty but we will forgive for the sake of the retailer and mother the possibility of integration into the WordPress© CMS. Particular respect is attached to the expressiveness of the generated movement.

The system was developed on the basis of deep neural networks and has the ability to integrate into the world's most popular content management system WordPress® and one system Google© Cloud Platform®.

2. *Methods and Materials*

The international standard is a document of the W3 Web Content Accessibility Guidelines (WCAG) 2.0 consortium [4].

DOI: 10.20535/2523-4455.mea.255961

The main provisions of the document are that the user should be able to:

- Perceive the components of the user interface and information in such a way that he was able to perceive them;
- Operate all components of the user interface and navigation;
- Clearly understand the structure of the user interface and the information presented;
- Use a variety of web browsers, including special browsers for people with disabilities.

A. CMS WordPress© Review

WordPress© is an open source content management system (CMS). The most popular way to create a website at the moment is CMS. CMS (Content Management System) is a system for creating and managing a site. Simply put, it is a program with a user-friendly interface for creating a web resource. You can use code in it, but it is not necessary. Thanks to this technology, anyone can create a page [5].

According to W3techs [4], WP is used by 64.7% of all CMS-built websites. This is 41.1% of all existing sites in the world. The New York Times and Forbes use this platform for their blogs. WordPress gained such popularity for its user-friendly interface and great features [6].

B. Natural language processing

Natural language processing (NLP) is an area at the intersection of computer science, artificial intelligence, and linguistics. The goal is to process and "understand" natural language to translate text and answer questions.

With the development of voice interfaces and chatbots, NLP has become one of the most important artificial intelligence technologies. But full understanding and reproduction of the meaning of language is an extremely difficult task, as human language has features:

Human language is a specially designed system for transmitting the meaning of what is said or written. This is not just an exogenous signal, but a conscious transmission of information. In addition, the language is encoded so that even young children can learn it quickly [7].

Human language is a discrete, symbolic or categorical signaling system that has reliability.

Categorical language symbols are encoded as signals for communication through several channels: sound, gestures, writing, images, and so on. Thus the language is capable to be expressed in any way.

C. Deep Learning in NLP

A significant part of NLP technology is due to deep learning, a field of machine learning, which began to gain momentum only at the beginning of this decade for the following reasons [6]:

- Accumulated large amounts of training data;
- Developed computing power: multicore CPU and GPU;

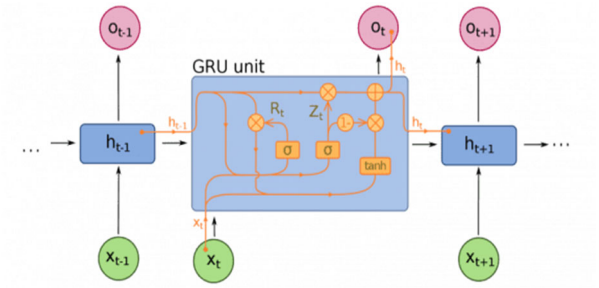

Fig. 1 Gate Reset Unit Algorithm

- New models and algorithms with advanced features and improved performance, with flexible learning on intermediate concepts;
- There are teaching methods using context, new methods of regularization and optimization [2].

Most machine learning techniques work well through human-developed representations of data and inputs, as well as optimizing weights to improve the final prophecy [9].

In deep learning, the algorithm tries to automatically extract the best features or representations from the raw input (Fig. 1). Hand-crafted traits are often too specialized, incomplete, and time-consuming to create and approve. On the contrary, the signs found by deep learning are easily adapted [10].

Deep learning offers a flexible, versatile, and one-taught framework for presenting the world in both visual and linguistic information. Initially, this led to breakthroughs in the areas of speech recognition and computer vision. These models are often learned using a single common algorithm and do not require traditional construction of features for a specific task [11].

D. Neural network machine translation

Approach to translation modeling using the Recurrent Neural Network (RNN). RNN is a neural boundary with dependence on previous states that has connections between passages. Neurons receive information from the previous layers, as well as from themselves in the previous stage. This means that the order in which data is input and the network is trained is important: the result of the Donald-Trump filing does not coincide with the Trump-Donald filing result.

The standard model of neural-machine translation is an end-to-end neural network, where the original sentence is encoded by an RNN called an encoder, and the target word is provided by another RNN called a decoder. The encoder "reads" the original sentence at the rate of one character per unit time, then combines the original sentence in the last hidden layer. The decoder uses error reverse propagation to study this union and returns the translated version. Surprisingly, on the periphery of research activity in 2014, neuro-machine translation became the standard of machine translation in 2016 [12].

Long / short term memory networks (LSTMs) try to deal with the problem of the disappearance gradient by entering gates and entering a memory cell. Each neuron is a memory center with three gates: input, output, and

DOI: 10.20535/2523-4455.mea.255961

forget. These shutters serve as security guards for information, allowing or prohibiting its flow.

- The input gate determines how much information from the previous layer will be stored in this cell;
- The output gate works at the other end and determines which part of the next layer learns about the state of the current cell.
- The gate of forgetting controls the extent to which the value is stored in memory: if a new chapter begins when studying a book, sometimes it becomes necessary for the neural network to forget some words from the previous chapter.

The reset gate functions are similar to the forget gate in LSTM, but the location is different. GRUs always transmit their full state without an output shutter. Often these shutters function like LSTM, however, the difference is that the GRU shutter is faster and easier to operate (but also less interpreted). In practice, they seek to neutralize each other, as they need a large neural network to restore expressiveness (expressiveness), which nullifies the gains in the result. But in cases where no extra expression is required, GRUs show a better result than LSTM [5].

Through the use and training of deep neural networks, the speech synthesis was adapted to the Ukrainian language and a package solution was developed (a plug-in archive is attached to the thesis) to integrate such a speech synthesis system into websites based on WordPress© content management system. at www.voice.uttermouse.com.

For the convenience of the visually impaired, the site has integrated voice chat, which was developed using Google© Cloud Platform and Google© Dialogflow and Google© Speech to Text API and adapted to work with the site using PHP.

E. Query Processing

The program is responsible for starting the web server and processing requests. you can use any web structure available for the development language. *The App* Engine runs multiple instances of the program, and each instance has its own web server *to process* requests. Any request can be redirected to any instance, so consecutive requests from the same user are not necessarily sent to the same instance. An instance can process multiple requests at the same time. The number of instances can be automatically adjusted as traffic changes.

The following example contains JavaScript code to start the server and respond to all GET requests from web clients to the root path. Importantly, in the last lines, the code provides the listening server with a port defined by the process.env.PORT variable. This is an environment variable set by App Engine - if the server does not listen to this port, it will not be able to receive requests. The incoming HTTP request includes HTTP headers sent by the client. For security reasons, *some have been removed,* headers are disinfected or supplemented by intermediate proxy servers before they reach the program. There are restrictions that apply to generated responses, and the response can be changed before it is returned to the customer.

Disable buffering By default, all App Engine responses are buffered in 64 container blocks. In some cases, it may be appropriate to disable buffering and pass bytes directly to the client. It is usually best to use suspended GETs or server events (SSEs). To disable buffering, you can set the X-Accel-Buffering response header to no. HTTPS connection For security reasons, all applications should encourage customers to connect to https. You can use the Strict-Transport-Security header to instruct the browser to prefer https over http to a specific page or the entire domain.

F. Overview of similar solutions

Amazon Polly Wordpress Plugin.

This plugin is developed on the basis of a deep neural network using the power of AWS.

Its advantage is that it provides a quick opportunity for voice acting due to the possibility of hosting servers in many countries (unfortunately, it is not represented in Ukraine).

The downside is that it does not integrate with the Kiki or WordPress editor and only provides support for shortcodes. Does not speak Ukrainian.

Speaker WordPress Plugin

The plug-in is developed using the same technologies as presented in the work, excluding the Scala language.

On the plus side, the plugin uses similar technologies, which allows you to independently train several neural networks.

The downside is that it does not integrate with the Kiki or WordPress editor and only provides support for shortcodes. Does not have Ukrainian language but can support.

Voice WordPress Plugin

The plug-in is developed using the same technologies as presented in the work, excluding the Scala language.

On the plus side, the plugin uses similar technologies, which allows you to independently train several neural networks. The plugin also has integration with the standard WordPress editor and integration with the popular WordPress editor Elementor.

The downside is that it does not integrate with all popular WordPress editors.

Responsive Voice Text-to-Speech

The plugin is developed on the basis of deep neural networks and Google Cloud server, has integration with popular WordPress editors, but cannot use semantic markup.

3. RESULTS

A. WordPress© plugin Creation:

Two types of hooks are used in WordPress©:

- Actions - allow you to add or modify WordPress© functionality.
- Filters - allow you to change the data.

DOI: 10.20535/2523-4455.mea.255961

Hooks are needed not only for plugin developers, but also for those who will use the plugin. Hooks are used everywhere: in the very core of WordPress©, in plugins and themes. It is the hooks that make WordPress so flexible. The plugin has three functions:

- register_activation_hook () - registers the function that will work when activating the plugin. Used to add plug-in settings, etc.
- register_deactivation_hook () - registers the function that should be started after deactivation of the plug-in. Used to delete temporary plugin data.
- register_uninstall_hook () - registers the function called when removing the plugin.

B. Configuration of Speech Signal Transmission and processing

In order for the neural network to be able to process the audio signal and be able to encode / decode it, the following algorithm was developed:

```
/ **<code> .Google.cloud.texttospeech.v1.AudioEncoding audio_encoding = 1 [(.Google.api.field_behavior) = REQUIRED]; </code>
   * @return int * /
  public function getAudioEncoding ()
  {return $ this-> audio_encoding;}
  / **<code> .Google.cloud.texttospeech.v1.AudioEncoding audio_encoding = 1 [(.Google.api.field_behavior) = REQUIRED]; </code>
   * @param int $ var
   * @return $ this
   * /
  public function setAudioEncoding ($ var)
  {
GPBUtil :: checkEnum ($ var, \ Google \ Cloud \ TextToSpeech \ V1 \ AudioEncoding :: class);
     $ this-> audio_encoding = $ var;
     return $ this;
  }
 <code> double speaking_rate = 2 [(.Google.api.field_behavior) = INPUT_ONLY, (.Google.api.field_behavior) = OPTIONAL]; </code>
   * @return float
   * /
  public function getSpeakingRate ()
  {      return $ this-> speaking_rate;}
  / **
  <code> double speaking_rate = 2 [(.Google.api.field_behavior) = INPUT_ONLY, (.Google.api.field_behavior) = OPTIONAL]; </code>
   * @param float $ var
   * @return $ this
   * /
  public function setSpeakingRate ($ var)
  {  GPBUtil :: checkDouble ($ var);
     $ this-> speaking_rate = $ var;
     return $ this; }
```

C. The data obtained:

@type int $ audio_encoding - Required. Audio byte stream format.

@type float $ speaking_rate - Speech speed in the range [0.25, 4.0]. 1 is the normal speed supported by a certain voice, 2.0 is twice as high, and 0.5 is twice as low. If unset (0.0), the default setting is 1.0. Any other values

0.25 or> 4.0 will return an error.

@type float $ pitch - Speech height, in the range [-20.0, 20.0]. 20 means an increase of 20 semitones from the initial pitch. -20 means a decrease of 20 semitones of the original pitch.

@type float $ volume_gain_db - Volume boost (in dB) from normal native volume, supported by a specific voice, in the range [-96.0, 16.0]. If not set or set to 0.0 (dB), it will play at the normal amplitude of the native signal. And the value of -6.0 (dB) will be reproduced at about half the amplitude of the native signal. The value will be +6.0 (dB).

@type int $ sample_rate_hertz - Synthesis sampling rate (in hertz) for this audio. When it is

@type string [] | \ Google \ Protobuf \ Internal \ RepeatedField $ effects_profile_id - an identifier that selects "audio effect" profiles that are applied to (after synthesized) text before broadcast.

DOI: 10.20535/2523-4455.mea.255961

The effects are applied in the order in which they are presented at https://cloud.Google.com/text-to-speech/docs/audio-profiles.

D. Decoding configuration.

<? php<code> Google.cloud.texttospeech.v1.SynthesizeSpeechRequest </code> {
/ **
The synthesizer requires either plain text or SSML as input.
<code> .Google.cloud.texttospeech.v1.SynthesisInput input = 1 [(.Google.api.field_behavior) = REQUIRED]; </code>
* / private $ input = zero; / **
* Voice synthesized audio.
* Generated from protobuf field <code> .Google.cloud.texttospeech.v1.VoiceSelectionParams voice = 2 [(.Google.api.field_behavior) = REQUIRED]; </code>
* /
private $ voice = zero;
/ *** Synthesized audio configuration.*
* Generated from protobuf field <code> .Google.cloud.texttospeech.v1.AudioConfig audio_config = 3 [(.Google.api.field_behavior) = REQUIRED]; </code>* /
private $ audio_config = null;/ **
* @param array $ data {
* Not necessarily. Data to populate the Message object.
@type \ Google \ Cloud \ TextToSpeech \ V1 \ SynthesisInput $ input
The synthesizer requires either plain text or SSML as input.
* @type \ Google \ Cloud \ TextToSpeech \ V1 \ VoiceSelectionParams $ voice
Synthesized audio voice required.
* @type \ Google \ Cloud \ TextToSpeech \ V1 \ AudioConfig $ audio_config
*}/ **
* Required. The synthesizer requires either plain text or SSML as input.
** Generated from protobuf <code> .Google.cloud.texttospeech.v1.SynthesisInput input = 1 [(.Google.api.field_behavior) = REQUIRED]; </code>
* @return \ Google \ Cloud \ TextToSpeech \ V1 \ SynthesisInput
* /

CONCLUSION

In the process, the project idea was described, which is to develop a batch solution for CMS WordPress© for voice content on websites [Fig.2]. Development of a user-friendly interface and installation system will help to enable the visually impaired to consume content on sites. The project was hastily developed using Google Cloud technologies, Kubernates and Java Script, PHP and Scala programming languages and applied on the websites http://ames.kpi.ua/ and voice.uttermouse.com/. Also, for successful work, a deep neural network was created and trained to reproduce Ukrainian speech, the neural network was trained for a year and a half using programmable tasks and speech samples.

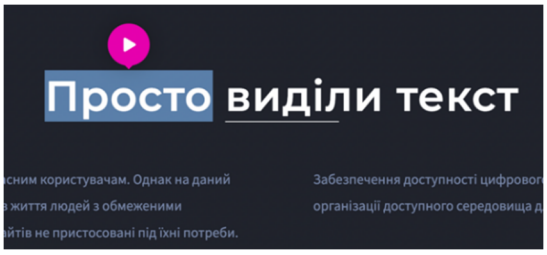

Fig. 2 Screenshot of the site with integration.

REFERENCES

[1]. Rabiner, L. Juang, B. H., *Fundamentals of Speech Recognition*, San Carlos, USA: Prentice-Hall International, Inc, 1993.

[2]. Hinton Geoffrey, Deng Li, Yu Dong, Dahl George, Mohamed Abdel-rahman, Jaitly Navdeep, Senior Andrew, Vanhoucke Vincent, Nguyen Patrick, Sainath Tara, Kingsbury Brian, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012. **DOI:** 10.1109/MSP.2012.2205597

[3]. W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Attend and Spell”, 2015. **URL:** https://arxiv.org/pdf/1508.01211.pdf

[4]. Prabhavalkar, R., Rao, K., Sainath, T.N., Li, B., Johnson, L., Jaitly, N., “A Comparison of Sequence-to-Sequence Models for Speech Recognition” in *Proc. Interspeech 2017*, pp. 939-943. DOI: 10.21437/Interspeech.2017-233

[5]. R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C. Chiu, A. Kannan, *Minimum word error rate training for attention-based sequence-to-sequence models*, 2017. **URL:** https://arxiv.org/pdf/1712.01818.pdf

[6]. C. Chelba, *Large Scale Language Modeling in Automatic Speech Recognition* **URL:** https://cloud.Google.com/dialogflow/docs/quick/setup

[7]. developers.Google.com, *Create a Project and Dialogflow Agent* **URL:** https://developers.Google.com/assistant/actions/dialogflow/project-agent.

[8]. cloud.Google.com , Dialogflow Documentation **URL:** https://cloud.Google.com/dialogflow/docs.

[9]. cloud.Google.com, Quickstart: Build an agent **URL:** https://cloud.Google.com/dialogflow/docs/quick/build-agent.

[10]. Statcounter Global Stats 2020
URL: https://gs.statcounter.com/

[11]. Marr, Bernard. How Artificial Intelligence IS Making Chatbots Better For Business. **URL:** https://www.forbes.com/sites/bernard-marr/2018/05/18/how-artificialintelligence-ismaking-chatbots-better-for-businesses/#69638bae4e72.

[12]. M. T. Mutiwokuziva, M. W. Chanda, P. Kadebu, A. Mukwazvure, and T. T. Gotora, "A neural-network based chat bot", in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, 2017, pp. 212–217. **DOI:** 10.1109/CESYS.2017.8321268

Надійшла до редакції 12 лютого 2022 року
Прийнята до друку 27 квітня 2022 рооку

DOI: 10.20535/2523-4455.mea.255961

УДК 621.3

Розробка системи і інтерфейсу синтезу мовлення українською мовою для сайтів

Осадчук О. Р., ORCID 0000-0003-4934-2565
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського" ROR 00syn5v21
Київ, Україна

***Анотація*—У статті описано систему синтезу та озвучування контенту на сайтах українською мовою, призначену для спрощення споживання контенту для людей з вадами зору, яка відрізняється простотою інтеграції в найпопулярнішу систему управління контентом на сайтах, а саме CMC Wordpress.**

Наразі люди з вадами зору дуже обмежені у використанні Інтернет-продуктів, оскільки більшість веб-ресурсів не адаптовані до їхніх потреб. Сучасні інформаційні технології дозволяють таким людям отримувати інформацію поряд із здоровою завдяки рішенням, розробленим вченими та інженерами з різних країн.

Текстова інформація може бути доставлена людям з вадами зору за допомогою лупи або шляхом програмного збільшення шрифту, незрячим – озвучуванням тексту за допомогою комп'ютерних програм або відображенням текстів на екрані Брайля монітора. Це ефективне рішення, але вибір методів відтворення такої інформації мають повністю забезпечувати самі люди з обмеженими можливостями, що є значною проблемою через значні витрати часу на споживання інформації.

Щоб полегшити сприйняття людей із вадами зору під час використання веб-сайтів, розроблено міжнародний стандарт для веб-майстрів – посібник із доступності веб-контенту під назвою W3 Web Content Accessibility Guidelines (WCAG) Consortium 2.0. Стандарт детально описує вимоги людей з вадами зору, які рекомендується виконувати, щоб вони могли без проблем переглядати веб-сайт. Основні положення WCAG 2.0 описуватимуть параметри та алгоритми масштабування, кластеризації та поділу інформації програмами для людей із вадами зору та нададуть рекомендації щодо написання коду веб-сайту.

Однак для реалізації таких рекомендацій веб-майстрам необхідно вивчити нові принципи та алгоритми програмування та використовувати додаткові інструменти розробки. Це часто буває складно, вимагає додаткової підготовки, що тягне за собою невиконання веб-майстрами таких вимог. Така система була розроблена на базі глибинних нейронних мереж та має можливість інтегруватися в найпопулярнішу в світі систему управління контентом веб сайтів WordPress® і однієї системи Google© Cloud Platform®.

За допомогою використання та навчання глибинних нейронних мереж було адаптовано процедуру синтезу мовлення під українську мову, а також розроблено пакетне рішення (архів плагіну додається до дипломної роботи) для інтеграції такої системи синтезу мовлення в сайти на базі системи управління контентом WordPress©, практична частина роботи реалізована на веб сайті www.voice.uttermouse.com.

Для подальшої зручності користування слабозорим на сайт був інтегрований голосовий чат, який розроблено за допомоги сервісів Google© Cloud Platform та Google© Dialogflow та Google© Speech to Text API і адаптований для роботи з веб сайтом за допомогою мови PHP.

Ключові слова — синтез мовлення; підтримка слабкого зору; алгоритм розпізнавання; обробка природної мови; нейронна мережа; розуміння природної мови; веб-інтеграція; CMS Wordpress.