РАЗДЕЛ II. ПРИКЛАДНАЯ ЛИНГВИСТИКА
SECTION II. APPLIED LINGUISTICS

UDC 81`33 **DOI: 10.18413/2313-8912-2022-8-4-0-6**

Andrey R. Biktimirov[1]
Dmitry Yu. Gruzdev[2]

Boosting Speech-to-Text software potential

[1] Military University
14 B. Sadovaya St., Moscow, 115432, Russia
E-mail: *andybikt@yandex.ru*

[2] Military University
14 B. Sadovaya St., Moscow, 115432, Russia
E-mail: *gru@inbox.ru*

Received 19 July 2022; accepted 12 December 2022; published 30 December 2022

Abstract. The article focuses on finding ways of boosting efficiency and accuracy of Speech-to-Text (STT)-powered input. The effort is triggered by the growing popularity of the software among professional translators, which is in line with the general trend of abandoning typing in favor of speech-to-text applications. Insisting that better effectiveness of such programs is contingent on their accuracy, the researchers analyze major factors, both linguistic and technical in nature, affecting the computer-assisted speech transcribing quality. This leads to an experiment, putting the hypothesis to a test. Based on numerical and performance data, errors and their breakdown into categories in an attempt to figure out their origins, it dwells on various approaches to dictation in a combination with several hardware options and configurations. These pave the way for recommendations on the improvement of STT performance based on the Dragon software. The authors arrive at a conclusion that it is possible to boost the STT accuracy up to 99 percent by adjusting the program profile to accommodate phonetic features of the speaker with due consideration of his accent, adding to the dictionary the most complex and rare vocabulary beforehand, and fine-tuning input hardware. Other noteworthy results include ways to overcome the most complex transcribing challenges, i.e. proper names, placenames, abbreviations, etc.

Keywords: Transcribing; Voice recognition; STT software; Dictation efficiency; Voice properties; Phonetic properties

How to cite: Biktimirov, A. R. and Gruzdev, D. Yu. (2022). Boosting Speech-to-Text software potential, *Research Result. Theoretical and Applied Linguistics*, 8 (4), 72-89. DOI: 10.18413/2313-8912-2022-8-4-0-6

УДК 81`33 DOI: 10.18413/2313-8912-2022-8-4-0-6

Биктимиров А. Р.[1]
Груздев Д. Ю.[2]

Способы повышения эффективности работы программы транскрибации речи

[1] Военный университет
ул. Б. Садовая, 14, Москва, 115432, Россия
E-mail: andybikt@yandex.ru

[2] Военный университет
ул. Б. Садовая, 14, Москва, 115432, Россия
E-mail: gru@inbox.ru

*Статья поступила 19 июля 2022 г.; принята 12 декабря 2022 г.;
опубликована 30 декабря 2022 г.*

Аннотация. Научная статья посвящена поиску способов повышения точности голосового набора текста с помощью программ транскрибации речи. Актуальность исследования обосновывается ростом популярности программного обеспечения (ПО) данного класса среди профессиональных письменных переводчиков, что обуславливает наметившуюся тенденцию перехода от ручного набора текста перевода к диктовке с преобразованием звучащей речи в текст. Авторы отмечают, что повышение продуктивности набора текста через диктовку зависит от точности работы программы распознавания речи. В работе анализируются основные факторы лингвистического и программно-технического характера, оказывающие наибольшее влияние на эффективность преобразования речи в текст компьютером. Для проверки выдвинутых предложений проводится эксперимент, в рамках которого анализируется количество ошибок транскрибации и причины их возникновения при использовании различных видов аппаратного обеспечения и способов диктовки. В результате исследования выявляются и предлагаются пути оптимизации работы с программой транскрибации речи на примере ПО Dragon. Авторы приходят к выводу о возможности повышения точности распознавания речи до 99% путем калибровки профиля программы под фонетические особенности речи с учетом акцента, пополнения пользовательского словаря наиболее сложной и редкоупотребимой лексикой до ввода текста, настройки технических средств голосового ввода. К другим значимым результатам следует отнести предложенные способы преодоления наиболее сложных трудностей транскрибации, таких как имена собственные, топонимы, аббревиатуры и сокращения.

Ключевые слова: Транскрибация речи; Голосовой ввод текста; Программа преобразования речи в текст; Эффективность диктовки; Свойства речи; Фонетические особенности речи

Информация для цитирования: Биктимиров А. Р., Груздев Д. Ю. Способы повышения эффективности работы программы транскрибации речи // Научный результат. Вопросы теоретической и прикладной лингвистики. 2022. Т. 8. № 4. С. 72-89. DOI: 10.18413/2313-8912-2022-8-4-0-6

Introduction

Automatic voice recognition (AVR) and speech-to-text (STT) software have been gaining popularity exponentially in numerous spheres ever since the introduction of deep learning approaches in the 2000s (Kan et al, 2018). The latter has become an effective alternative to typing in translation, education, business, routine paperwork, and even medicine and police investigations (Ogunshile, Phung & Ramachandran, 2021; Trabelsi, 2022; Kurzekar et al., 2014). Some visionaries in the field even seek ways of integrating the solution in military technologies, in particular C2 systems, no matter how noisy and incoherent communications can get on the battlefield (Cornaggia-Urrigshardt et al., 2022).

As handwriting was once abandoned in favor of keyboard, voice input will eventually replace typing. It is already happening for several reasons. First, dictating is faster than typing. Statistically, the rate of normal speech is 120–150 words per minute, while the speed of typing of an average person is around 100 characters per minute (Ibid, 2021). The fact that speech evolves naturally is another argument in favor of the state-of-the-art technology. Second, the advanced stage of the software development provides high quality and desired accuracy even on mobile devices, like smartphones. Many people already take advantage of it to input and send short texts in messengers (Kumar, Gupta & Sapra, 2021; Ogunshile, Phung & Ramachandran, 2021). New projects in this field has a potential not only to step up the effectiveness, but even expand the footprint of the technology application. Programmers have zeroed in on more complex tasks, namely streaming STT for TV-news to accommodate the needs of hearing-impaired citizens or provide online subtitles (Jorge, Giménez, Baquero-Arnal et al., 2021; Perero-Codosero, Juan, Fernando et al., 2022, Kuzmin & Ivanov, 2021). Needless to say, the uncontrolled noise environment inherent in such fields presents a challenge (Montegro et al., 2021). Third, ceaseless projects undertaken to expand the STT footprint to more languages bring closer the moment of abandoning pens and keyboards in favor of voice inputs as a primary text-making tool (Brucal et al., 2021; Messaoudi et al., 2021).

This said, it is still not natural to talk to a device, demonstrating no signs whatsoever of understanding. The fact makes many new users put unnatural emphasis on pronunciation and articulation in the process of dictation. It just might not make much difference, since computers can distinguish sounds which are not pertinent to a particular language, thus not registered by the native speakers[1]. However, stripped of other means of elaboration of the speech, i.e. gesticulations, eye contact, etc., those trying to master the new technology fall back on the need to pronounce words far more accurately, since without this grammar and vocabulary alone, no matter how good they are, will not provide understanding by the recipient of the message (Stubna, 2020). Thus the transition to full-fledged use of STT in routine operations calls for adapting to, for example, the need to spell punctuation marks, which is not natural for human beings, and voice commands to navigate in computer programs. In some areas much more has to be done in terms of integration. In translation it has to do with the major task of conveying the message and de facto replacement of translation with sight interpretation (Gruzdev, Gruzdeva & Makarenko, 2019).

These obstacles do not stop researchers from pursuing far-sighted ideas of mating STT and machine translation (Ma, Nguyen & Ma, 2022). However, as long as translation is still a Human-In-The-Loop trade, the primary focus of the article is on the STT aspect. Since the issue belongs to applied linguistics, it is natural that this field has accumulated considerable practical data, now needed to be sorted out and studied. Modern translators

[1] Harbeck, J. (2015). The subtle sounds that English speakers have trouble catching, *The Week*, retrieved from https://theweek.com/articles/569137/subtle-sounds-that-english-speakers-have-trouble-catching (Accessed 14 November 2022).

harnessed the technology several years ago, which is attested by discussions on special forums and the fact that developers of translation software, e.g. computer assisted translation, a.k.a. CAT, advertise the option of mating their products with STT input methods.

Goal

Currently, the number of different types of speech-to-text (STT) software on the market of digital products is growing, as well as their quality and popularity among specialists. Due to the application of the latest technologies there is a steady growth of the productivity of this class of programs and the increase of accuracy up to 99 percent.[2] This presents a stark contrast to the first attempts at making "Audrey" of the Bell Labs understand numbers in 1952 and some simple words some years later (Deng et al., 2013).

Today, Nuance's research data shows a three-fold increase in typing speed with STT programs.[3] As part of a research on increasing the productivity of translation through dictation, it was concluded that the greatest loss of time comes from post-editing the target text. For this reason, achieving maximum recognition accuracy is considered to be one of the most important aspects of the software application. In this respect, the goal of the paper is to identify the main factors that reduce the accuracy of speech recognition and propose ways to eliminate them.

Materials and methods

In previous studies a comparative analysis of the characteristics and features of the most accessible and popular STT software was conducted. Given the aggregate total of advantages, the choice was made in favor of Dragon by Nuance.

Firstly, from a technical point of view, this software does not have special hardware requirements, is compatible with most operating systems and can run off-line. Therefore, such factors as insufficient performance or quality of hardware (computer, microphone, voice recorder, etc.), as well as low-speed Internet connection have been excluded.

Second, in terms of functionality, the software comes with such features as a customizable user profile, audio file processing in various formats, an updatable user dictionary, and a voice control and editing interface. This allows the user to improve the accuracy of speech transcribing by individually adjusting the profile (Gruzdev and Biktimirov, 2022).

In order to assess the degree of influence of various factors on the STT quality, a text dictation experiment was conducted to encompass various equipment configurations and operating modes. It was decided to use the 2015 Dragon program produced by Nuance. The translation process was excluded from the experiment to avoid additional cognitive load, preventing a full-fledged assessment of software performance and making the effort to develop recommendations on improving the recognition quality a more complex endeavor. For the dictation, an original English text, 2,735 characters long, was chosen.[4] Since the efficiency of the transcribing process depends on the ability of hardware and software to perceive and decode speech, more problems are expected with a non-standard vocabulary, which does not constitute a part of the common used lexicon (Nugraha & Dewanti, 2022). To cover this issue, a complicating factor in the form of a large number of placenames and proper names, all foreign to the English culture, was introduced.

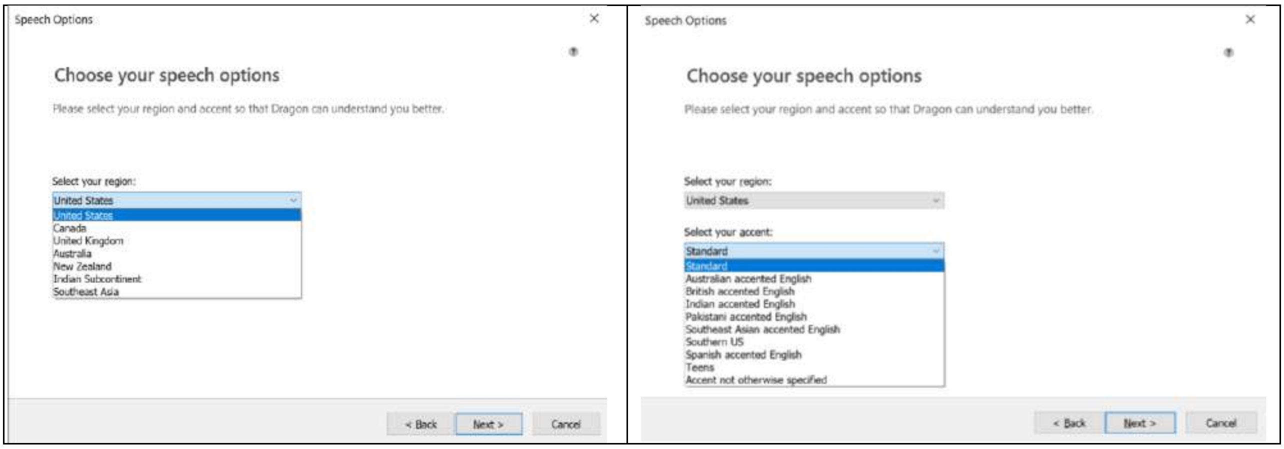
Before the experiment, the user profile was calibrated to the phonetic features of the speaker's speech. The main part of the

[2] Nuance (2022), retrieved from https://www.nuance.com/ (Accessed 04 June 2022).
[3] Ibid.

[4] Hsu, S. S. (2021). Judge rules Afghan militant has been held in Guantánamo illegally, in what lawyers say is the first such ruling in 10 years, *The Washington Post*, retrieved from https://www.washingtonpost.com/local/legal-issues/guantanamo-bay-detainee-held-illegally/2021/10/21/1a54245e-31c2-11ec-9241-aad8e48f01ff_story.html (Accessed 25 February 2022).

procedure was reading of a passage. This was preceded by adjustments done to the extent of choosing the region and accent (see Figure 1). However, there are no suitable settings among the offered options, reflecting Russian-accent in the English language. Thus, it was decided to use the British variant of the English pronunciation.

Figure 1. Dragon speech settings
Рисунок 1. Выбор региона и акцента для калибровки программы «Дрэгон»



Speech Options

Choose your speech options

Please select your region and accent so that Dragon can understand you better.

Select your region: United States

Speech Options

Choose your speech options

Please select your region and accent so that Dragon can understand you better.

Select your region: United States

Select your accent: Standard

Standard
Australian accented English
British accented English
Indian accented English
Pakistani accented English
Southeast Asian accented English
Southern US
Spanish accented English
Teens
Accent not otherwise specified
< Back
Next >
Cancel

In order to assess the impact of several types of input hardware on the quality of the program it was decided to use a professional microphone, a digital voice recorder, and a smartphone. The text was dictated twice into the microphone, recorder and smartphone, after which the results were subjected to analysis of the influence of extraneous noise, dictation tempo and diction on the accuracy of speech recognition. First, the text was read to all of the indicated assets with an emphasis on clear pronunciation while disregarding speed. The procedure was then repeated at a faster rate of about 100 words per minute. No corrections or pauses were allowed.

Several factors led to the choice of the selected means of input. First, the microphone option replaces only one tool of a modern translator, namely the keyboard, preventing profound adjustments to the general translating procedures. Second, the intention to use a voice recorder is accounted by the need to test the functionality of recorded audio transcribing feature, which distinguishes Dragon from most of this software class. Third, in case of good performance, the smartphone as the main audio recording tool in hand today would be worth considering in the translation strategy, replacing typing on the keyboard with dictation.

When doing streaming STT, particular attention was paid to avoiding microphone-assisted corrections, as such attempts could result in significantly longer dictation time and complete loss of usefulness of the alternative input method. In all dictation options, voice commands for punctuation marks were not voiced, which provided grounds for testing the effectiveness of the automatic punctuation feature.

Statistical analysis

For a general picture of the software effectiveness, it is required to provide statistics first. The STT accuracy was calculated based on the ratio of mistakenly recognized words to the total length of the text. The results of the experiment are in Table 1.

Given the slight difference in speech recognition quality achieved at fast and slow dictation rates, it was decided to base further analysis on the variants with the lowest tempo readings in each of the three subgroups. The choice was made in view of the main goal of maximizing the performance of the translator using the alternative input method.

Table 1. STT performance, conditioned by the speech rate and input equipment
Таблица 1. Анализ результатов транскрибации в зависимости от темпа речи и применяемого типа аппаратного обеспечения

No. seq.		Test phase	Results			
			Characters, no spaces/words	Dictating/recording processing (min:sec)	Speech rate (word/ min)	Errors (%)
0		Original text	**2,735/550**	-	-	-
1	1	Mic-assisted, articulated	**2,665/547**	**6:46/-**	**81**	**40 (7%)**
	2	Mic-assisted, moderate speed	**2,659/550**	**6:32/-**	**83**	**38 (7%)**
2	1	Recorder, articulated	**2,705/554**	**5:54/0:22 (6:16)**	**94**	**28 (5%)**
	2	Recorder, moderate speed	**2,725/557**	**5:44/0:21 (6:05)**	**97**	**33 (6%)**
3	1	Smartphone, articulated	**2,615/546**	**5:54/0:22 (6:16)**	**94**	**34 (6%)**
	2	Smartphone, moderate speed	**2,705/558**	**5:44/0:21 (6:05)**	**97**	**31 (6%)**

The analysis of the texts revealed a pattern of errors in terms of where they appear in the text and their density. In order to confirm the finding, error plots were made for each text using the AntConc program (see Table 2).[5]

It follows from the plots that the error density is almost the same in all dictation options, mostly found in paragraphs with proper names, which are a particular issue for the software due to the poor standardization of spelling. The situation with names of foreign origin is more complex (see Appendix 1).

In the highlighted passages there are also errors of another nature, but they are in no way related to the presence of a large number of proper names. No accumulation of errors due to uncommon collocations has been observed. Correct pronunciation of the lexical units following a mistakenly transcribed word reduces the probability of distortions. The increased density of errors is accounted for by difficulties with the pronunciation of proper names, which requires better concentration, while the recovery of the speech organs and their return to normal functioning in conditions of reproduction of unfamiliar sounds does not occur immediately. As a result, some sounds which are not pertinent to the pronounced words, transpire, laying the ground for incorrect STT performance.

It comes from the comparison of the STT results that the microphone input quality is inferior to recorder- and smartphone-powered transcribing options. In some cases, the smartphone proved to be the most effective tool. After listening to the voice data, it was established that the microphone and recorder sound tracks are less stable; the sound progresses in waves, fades or increases sharply, after which it recovers to normal values. Such areas of the recordings coincide with the main zones of errors made during transcribing.

[5] Anthony, L. (2020). AntConc (3.5.9) [Computer Software]. Tokyo, Japan: Waseda University, retrieved from https://www.laurenceanthony.net/software (Accessed 15 February 2022).

Table 2. STT error plots
Таблица 2. Графический анализ распределения ошибок в текстах, набранных методом транскрибации

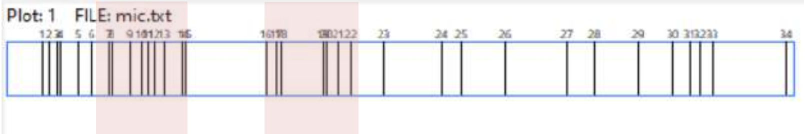
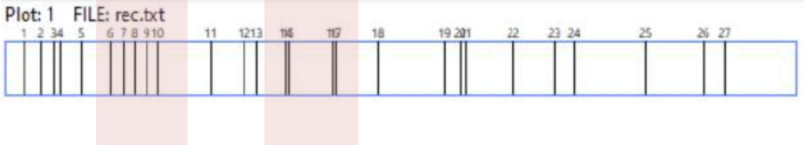
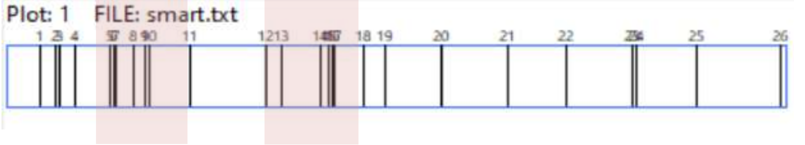
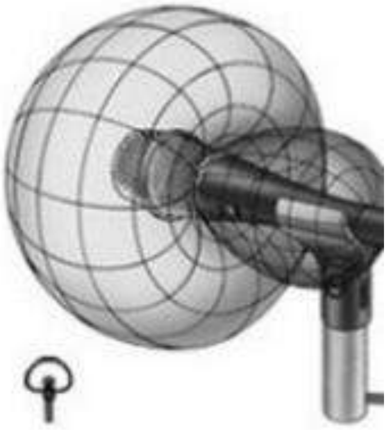
Error quantity	Error plot	Input method
Lapses: 7 Errors: 32		mic
Lapses: 7 Errors: 39		recorder
Lapses: 3 Errors: 31		smartphone
	Annex 1 Exerpt 1 Annex 1 Exerpt 2	

Figure 2. Hypercardioid directional pattern
Рисунок 2. Гиперкардиоидная диаграмма направленности



The studio microphone and the semi-professional voice recorder used in the experiment possess all properties and characteristics needed to perceive all changes in the voice flow. Both instruments are directional, i.e., they pick the signal in a narrow area, which makes them less sensitive to surrounding sounds (see Figure 2).[6] At the same time, the slightest jumps in tone, tempo, and strength, which were observed in segments of texts with numerous proper names, will also be registered, increasing the chances of STT errors.

[6] Mic direction patterns, retrieved from https://gs-corp.ru/articles/articles/Mic-direction-info/ (Accessed 11 May 2022).

The waveforms generated by Audacity allowed to establish the relationship between the unstable audio signal and the occurrence of errors in the text (see Table 3).[7] In the selected segments captured on the recorder, sharp spikes in signal strength were observed in the pronunciation of incorrectly transcribed words. Errors are registered not only in the words corresponding to the spikes in waveforms, but in neighboring lexical units as well.

At the same time, inadvertent voice sounds do not always lead to distortions in the typed text. In a number of cases, incorrectly started words followed by correct pronunciation or unintentional attempts to insert non-existent prepositions did not result in transcribing errors (see Figure 3). In the waveform of the passage "to comment" the inadvertently voiced preposition <of> was registered, which was not added by the program to the text. It should be noted that the trend is observed in cases where the power of the sound signal is insignificant. When the signal peaks above the strength threshold, the program begins to process the incoming information, which leads to the conversion of the erroneously pronounced sounds into written text.

Table 3. Recorder- and smartphone-based waveform of the phrase “classified opinions on”
Таблица 3. Волноформы фразы “classified opinions on”, сгенерированных из записей на диктофон и смартфон

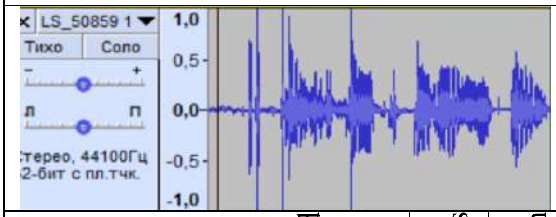
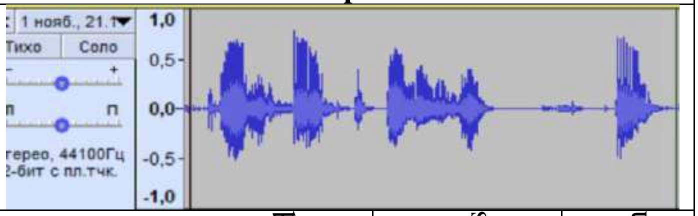
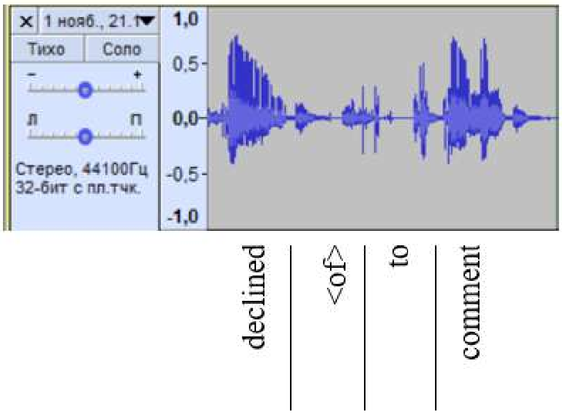
Recorder	Smartphone
 LS_50859 1 Тихо Соно Стерео, 44100Гц 32-бит с пл.тчк. 1,0 0,5 0,0 -0,5 -1,0	 1 нояб., 21.1 Тихо Соно Стерео, 44100Гц 32-бит с пл.тчк. 1,0 0,5 0,0 -0,5 -1,0
classified opinions on	classified opinions on
classify**(ied)** **convenience** on	classified opinions on

Figure 3. Smartphone-based waveform of the phrase “declined to comment”
Рисунок 3. Волноформа фразы “declined to comment”, сгенерированной из записей на смартфон



[7] Audacity Team (2021). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.0.0 retrieved March 17, 2021, retrieved from https://audacityteam.org/ (Accessed 5 April 2022).

Another disruptive factor introduced with the replacement of the familiar keyboard with voice input is the need to pronounce punctuation marks. Due to the natural replacement of punctuation by intonation in spoken language, the need to voice punctuation leads to inconveniences that distract from conveying the message and complicate the process of information reproduction (Kol´aˇr & Lamel, 2012). In this paper, the task of voicing punctuation marks was ignored by design only to evaluate the automatic punctuation feature, thus the errors of this nature were not considered in the overall statistics. The program managed to place dots and commas automatically with a moderate number of errors (see Appendix 1). Other punctuation marks, such as brackets, dashes, etc., were not implemented in the printed text. In view of this, it does not seem appropriate to consider this function as a complete automating punctuation solution in voice typing. However, the text becomes more readable, which speeds up post-editing (Kol´aˇr & Lamel, 2012). The use of the function is limited to recording transcribing.[8] Streaming input is not supported.

For a more in-depth study of STT errors, finding and identifying ways to eliminate them, it is necessary to perform their qualitative analysis in the selected passages in Appendix 1.

Qualitative analysis

STT errors are accounted for by the fact that the software in question is designed to decode the flow of spoken speech into written form. The underling factors range from linguistic features of the language to human factors. Errors are by no means chaotic. By categorizing the causes, it will be possible to eventually reduce the number. As a result, five categories were identified, namely: (1) homophones, (2) hesitation and delimitation, (3) incorrect pronunciation, (4) proper names, (5) other errors (see Table 4).

The first thing to remember is that STT software is a set of algorithms. Not able to understand to take into consideration all factors of spoken language, the software can never demonstrate 100% accuracy when decoding homophones: won/ [wʌn] - one, two/ [tuː] - to, their/ [ðeə] - there, court/ [kɔːt] - caught, counsel/[ˈkaʊnsəl] - council (see Appendix 1) (Pernarčić, 2019). It is these errors that make up the first category. Due to the nature of the English language, this group also includes nouns, sounding similarly in the plural and possessive forms. This said, it is worth noting the insignificant number of errors caused by phonetic ambiguity. Most cases were caused by the low frequency of use in the language and the lack of a clear contextual ties of the words to their immediate surrounding. For example, the numeral *two* and noun in the plural form *opinions* are wedged by the adjective *classified*, which was incorrectly decoded by the program as a verb in the infinitive form. As a result, the numeral turned into the particle *to*. The British National Corpus provides no matches of the sought collocation.

Parasitic speech sounds caused 1.5 times as many errors as were registered in the previous group. Loud sounds, for example, hm..., um..., etc, sharp and deep breaths and exhalations, as well as background sounds caused spontaneous insertions of articles, prepositions and pronouns in the printed text. The appearance of irrelevant text elements or the omission of deliberately pronounced units is accounted for by the sound threshold, i.e., a certain signal strength when the program begins to decode the input. Poorly articulated words with a volume lower than the main stream of speech leads to the fact that the program perceives the incoming information as noise. This tendency is evident when comparing waveforms of passages with indefinite articles (see Figures 4 and 5). In the first case, the article is clearly detected with sufficient signal strength, while in Figure 5 there is a signal, barely visible on the waveform, that has not been processed by the program.

[8] Nuance.

Table 4. STT errors breakdown
Таблица 4. Распределение ошибок транскрибации речи в зависимости от способа ввода

Input method	Errors						
	Homophones	Hesitation and delimitation	Accent-generated errors	Proper nouns	Other	TOTAL	Error rate, %
Mic	**4**	**7**	**4**	**19**	**4**	**38**	**7%**
Recorder	**3**	**8**	**4**	**18**	**0**	**33**	**6%**
Smartphone	**4**	**5**	**3**	**19**	**0**	**31**	**6%**

Figure 4. Smartphone-based waveform of the phrase “a former [Afghan militant]”
Рисунок 4. Волноформа фразы “a former [Afghan militant]”, сгенерированной из записей на смартфон

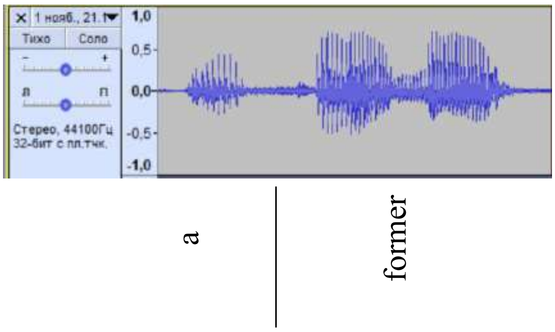
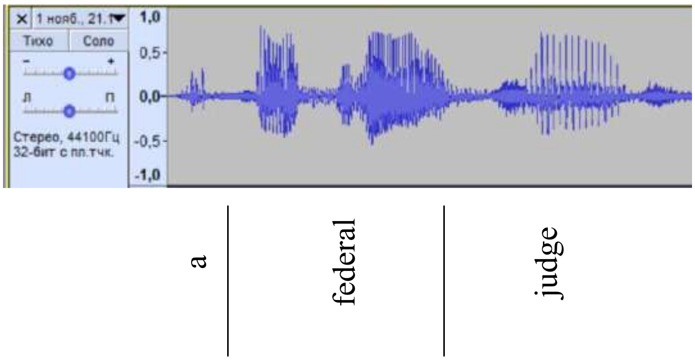
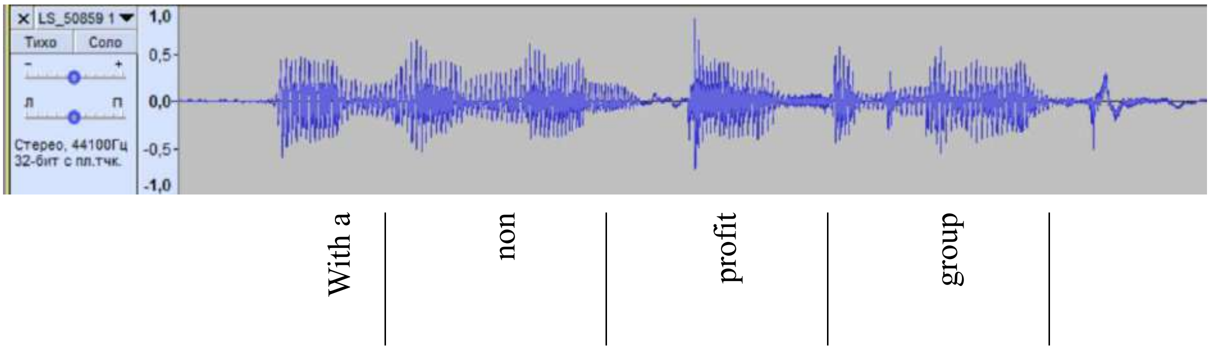


Figure 5. Smartphone-based waveform of the phrase “<a> federal judge”
Рисунок 5. Волноформа фразы “<a> federal judge”, сгенерированной из записей на смартфон



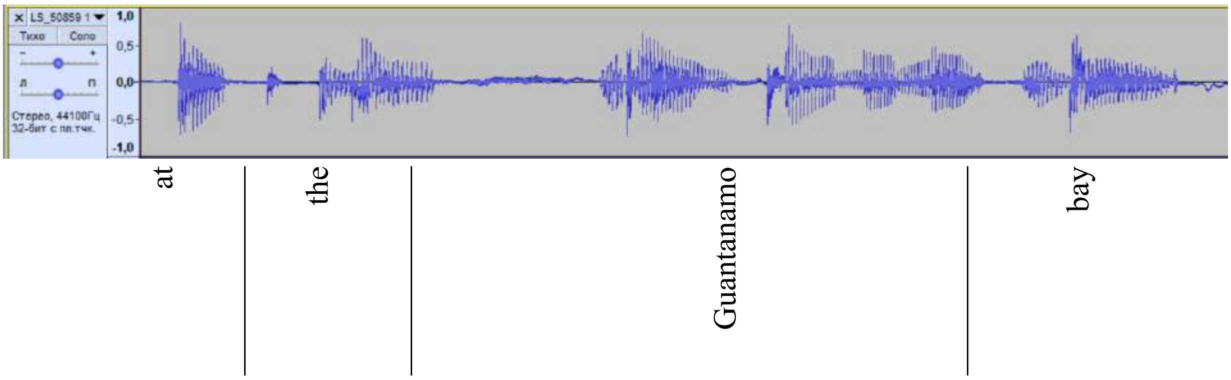
In addition to numerous hesitation pauses, the second category includes delimitation errors caused by incorrect definition of word boundaries by the software (Belitskaya, 2014). For example, as a result of a subtle pause between the preposition and article during the pronunciation of the phrase "...with a nonprofit group..." (pronounced [wɪð-ə---nɒn'prɒfɪt---gru:p]) the combination of the preposition *with* and article *a* triggered the substitution of *a* with *the* (see Figure 6).

Figure 6. Smartphone-based waveform of the phrase “with a nonprofit group”
Рисунок 6. Волноформа фразы “with a nonprofit group”, сгенерированной из записей на смартфон



It follows from the analysis of the recordings and waveforms of the sections of texts where incorrect transcribing of the definite article was registered that in most of the cases it was pronounced [ðə]. When the article assumed its strong phonetic form [ði:] such a problem was not observed (see Figure 7). Missing articles situations are due to the predominance of the weak form in the general speech.

Figure 7. Smartphone-based waveform of the phrase “at the Guantanamo Bay”
Рисунок 7. Волноформа фразы “ at the Guantanamo Bay”, сгенерированной из записей на смартфон

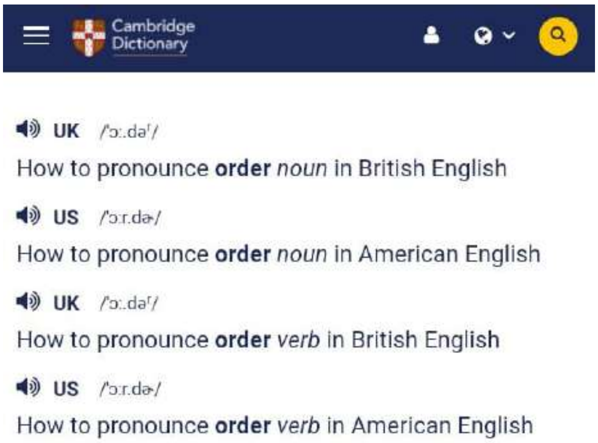


Hesitations are also a consequence of the natural desire to follow the graphical result on the PC monitor when doing STT into the microphone. As a result, the translator is distracted from the main task, and the delayed processing of the voice signal disrupts the dictation tempo, forcing the translator to stop in the middle of a sentence, which affects the integrity of the text. Thus, editing during translation should be done after finishing a meaningful text segment, such as a paragraph. On the other hand, increasing pace leads to loss of information. It will be more expedient to maintain an optimal speech rate, which allows articulating all the sounds.

Hesitation pauses between sentences or meaningful text parts can be eliminated by means of the press-to-talk (PTT) feature on the input device. The Dragon manufacturer has made this available in its standard microphones. The recorder also allows to adopt a similar strategy by using the PAUSE key.

Pronunciation, typical for the variant of the English language other than the one selected in the profile settings, also caused a number of errors. For example, the phrase *final order* is transcribed with significant errors, *final oral, final quarter*. This is accounted for by the sound [r] ([ˈfaɪn(ə)l ˈɔːrdər]), tended to be reduced in the British version ([ˈfaɪn(ə)l ˈɔːdə]) (see Figure 8). In view of the obvious trend, it was decided to group these STT inaccuracies into a separate category.

Figure 8. Pronunciation of the word *order* in the Cambridge Dictionary
Рисунок 8. Транскрипция английского слова *order* в Кебриджском словаре



UK /ˈɔː.dəʳ/
How to pronounce **order** *noun* in British English

US /ˈɔːr.dɚ/
How to pronounce **order** *noun* in American English

UK /ˈɔː.dəʳ/
How to pronounce **order** *verb* in British English

US /ˈɔːr.dɚ/
How to pronounce **order** *verb* in American English

It is worth noting the complex nature of individual cases in the category. Thus, when processing the word *necessitates*, voiced according to the American English standards [nəˈsesəˌtets] instead of the pronunciation adopted in British English [nɪˈsesɪteɪts], the program typed the noun *necessity* in one instance. In another iteration this led to a delimitation error, resulting in three words *and assists dates*. Both cases were detected in texts dictated at a slow rate. As the dictation speed increased, the number of errors in processing complex words pronounced in accordance with the phonetic norms of the other variant of English decreased (see Table 5).

Table 5. American accent-provoked errors
Таблица 5. Ошибки, вызванные произношением по нормам американского варианта английского языка

Original	Mic	Smartphone	Speech rate
Lawyers for those detainees say their continued detention despite having been cleared necessitates action by a court.	lawyers for those detainees ***see*** their continued detention despite having been clear**(ed necessitates)** action by ***the*** court	Lawyers for those detainees say ***they are*** conducted detention despite having been cleared necessitates action by a court	Moderate
	lawyers for those detainees ***see the*** continued detention despite having been cleared ***necessity*** action by ***the*** court	Lawyers for those detainees say ***there*** conducted detention despite having been cleared ***and assists dates*** action by a court	Low

The analysis of this category of errors suggests that it is possible to reduce their number by increasing the clarity of speech, correcting pronunciation in accordance with the phonetic norms of the selected language variant, articulating word endings. No less important is the correct calibration of the user profile, with due consideration of the accent.

Most of the aggregate total STT errors, incorrect spelling of proper names has been registered in each of the input options. There have been significantly fewer errors registered in proper names of European origin than Oriental or Asian names (Khan, Asadullah Haroon Gul, Amit P. Mehta). As a result, *Khan* became *hand, can, clan,* and *town*, and *Asadullah Haroon Gul* turned into *la blue girl*, *are circular her own gulf*, and *I supply her own gal*. More euphonious names for English speakers – Roman, Tara, Alex Brandon, Mark Meyer, Asad, Biden, etc. – were typed by the program correctly in most cases. One way to eliminate errors of this nature is to take advantage of Dragon's special feature, its open vocabulary. If a certain name is used frequently in a text, it is advisable to add it to the user database in advance. For single cases, a more effective and faster solution would be its omission or descriptive rendering based on the function performed by the person, his position or role in the described situation (see Table 6).

Table 6. Real-time STT dictation
Таблица 6. Голосовой ввод через микрофон в режиме реального времени

Proper names kept	Proper names replaced
earlier this month in a separate proceeding the US government determined that it was safe to transfer ***well*** who has never been charged with a crime out of Guantánamo he is among 13 men who have been recommended for transfer by the multi agency periodic review board PRB on the basis that they are not considered to pose a threat to US national security	Earlier this month in a separate proceeding the United States government determined that it was safe to transfer **the prisoner** who has never been charged with a crime out of Guantánamo. He is among 13 men who have been documented for transfer by the multiagency periodic review Board on the basis that they are not considered to pose a threat to US national security.

Ad-hoc acronyms and abbreviations, which are not uncommon especially in science and engineering, can also be subject to incorrect transcription due to their absence in the vocabulary (Chistova, 2021). In this case, potentially complex acronyms and abbreviations should be added to the profile database or replaced with a striking, out-of-the-context substitute word to be edited or changed to the necessary lexical unit through the auto-replacement function of the text editor.

A number of errors have not been assigned to any of the highlighted categories. Though no phonetic norms were broken, each of these turned up only once in one of the six STT variants: *Afghan refugee* [ˈæfgæn ˌrefjʊˈdʒiː] - *and come refugee*, *the opinion* [ðiː əˈpɪnjən] - *their pinion*, *classified opinion* [əˈpɪnɪənz] - *classified convenience*, *Biden's troop* [truːp] - *Biden's group*. However, the recordings unveiled unstable sound quality in most cases, which probably led to the errors. Considering their singular character, they have an insignificant impact on the overall statistics.

Most of the identified problems can be solved by approaching the issue from two general perspectives. First, the input devices and software should be adjusted. Secondly, it is necessary to maintain an optimal speech rate at which it is possible to ensure articulation of all elements of the text, including endings and service parts of speech, often weakened in speech due to their auxiliary functions in the sentence. An important role is assigned to signal clarity, free from hesitations and noises.

Corrections

To check the suggestions for improving the STT quality, another text was picked containing terminology and proper nouns. The equipment was adjusted and fine-tuned with all necessary preparations done beforehand.[9] First, the sensitivity of the voice input devices was reduced and a noise filter was enabled in the settings. Second, the most difficult and rarely used vocabulary was identified to include special terms, abbreviations, acronyms, company names and proper nouns and added to the user vocabulary.

When dictating, due considerations were given to the English variant, to which the profile was set, focusing on the proper pronunciation of articles, prepositions and word endings.

The milestone text was divided into three equal segments. The first segment was dictated through a microphone, while the second and third were recorded on the smartphone and recorder, respectively. The STT accuracy grew to 99 percent (see Table 7).

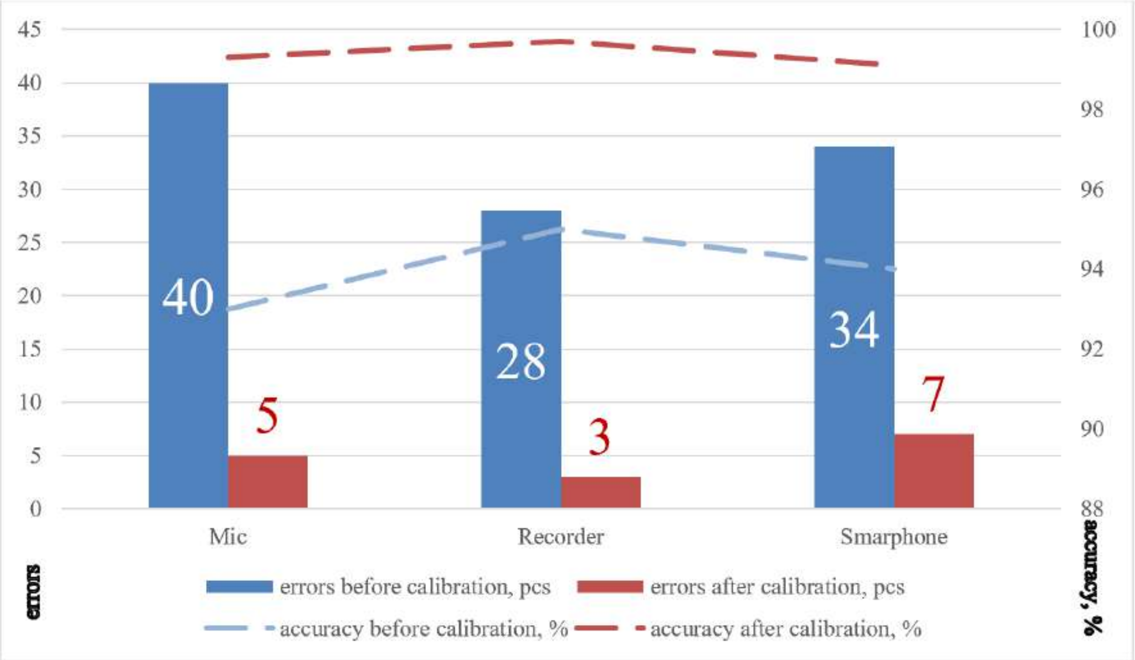
The recordings confirmed that all errors resulted from incorrect pronunciation, mostly due to ignoring vowels in stressed syllables. After fine-tuning of the equipment and vocabulary preparation the transcribing efficiency improved significantly, which has been established through a comparative analysis of the quantitative readings accumulated in both phases of the experiment (see Figure 9).

Table 7. Milestone STT statistics, based on lessons learned
Таблица 7. Статистика транскрибации речи при диктовке контрольного текста

Input method	Results		
	Words	Errors	Accuracy, %
Mic	672	5	99.3%
Smartphone	727	7	99.1%
Recorder	907	3	99.7%
TOTAL	2,306	15	99.4%

[9] Antonov, N. (2017). Let there be light… lidar, *Top War*, retrieved from https://en.topwar.ru/125104-da-budet-svet-lidara.html (Accessed 20 February 2022).

Figure 9. Setting- and calibration-based comparative analysis of STT accuracy
Рисунок 9. Сравнительный анализ эффективности транскрибации с учетом базовых настроек и калибровки



Conclusions

The research shortlists major linguistic and software factors that have the greatest impact on the STT accuracy. Based on the experimental data, recommendations were made to improve the efficiency of this class of software and boost its accuracy to 99 percent. Firstly, it is possible to use different input devices without incurring significant losses in the quality. However, the minimum number of errors was registered when a professional digital voice recorder was used, provided all hard- and software adjusted and fine-tuned. Transcribing a recording proved to be a faster option than streaming voice data into a microphone, as text visualization kept distracting from the main task, provoking unwanted hesitations and delimitation errors, acting as “speed bumps” to the tempo. Second, attention should be paid to maintaining an optimal speech rate that will allow all sounds to be articulated. Acceleration above the average tempo inevitably leads to more errors and omissions. In this regard, it is necessary to note a lower susceptibility of wired systems to interference and, by implications, losses of voice data. The use of wireless technology also provides good capturing of the audio stream, although inconsistent at times. On the other hand, an extremely slow rate will lead to an increased number of delimitation errors. Third, accuracy can also be boosted through the adjustment of input devices. Setting the sensitivity of the microphone or recorder correctly will reduce background noise. However, a significant drop in sensitivity can lead to the omission of short parts of speech when they are pronounced at a volume lower than that of the main stream. Some devices, such as smartphones, lack such adjustments and tuning options. Fourthly, the program is based on the analog principle of audio input processing, thus the avalanche-like generation of errors is an unlikely scenario. Nevertheless, the STT texts have pockets of inaccurately transcribed words and their combinations. As a rule, such strings of errors are triggered by a word, pronunciation of which is not typical for the phonetic system of the language. Destabilization of the vocal organs leads to jumps in the strength of the incoming signal and rate of speech, causing a detrimental effect on the quality of data registration. Fifth, the vocabulary with destabilizing potential includes long proper names, especially with complex phonetics, specialized abbreviations and acronyms. They account for more than 50 percent of all errors. Adding them to the user vocabulary in advance will step up the overall efficiency. For rare cases, it is advisable to replace phonetically complex units with a substitution word that will attract the attention at the editing stage, or implement them in the text descriptively, e.g., through their roles or functions assigned in the described situation. Sixth, if there are several variants of the language, their phonetic systems should not interfere with each other. This aspect should also be considered when performing the initial setting of the profile. Seventh, automatic punctuation is only available in the recording transcribing mode. Because of the functionality limited to two symbols, the dot and comma, the range of potential users of the tool is narrowed down to the beginner group, who will find it difficult to immediately adjust to the need to voice punctuation marks. With practice, the computer-assisted approach can be replaced with voice input.

References

Belitskaya, A. (2014). Roles of hesitation pauses in spontaneous speech, *Philology and literary studies*, 2, available at: https://philology.snauka.ru/2014/02/698 (Accessed 20 October 2022). *(In Russian)*

Brucal, S. G. E. et al. (2021). Filipino speech to text system using Convolutional Neural Network, *Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4)*, 176-181. DOI: 10.1109/WorldS451998.2021.9513991 *(In English)*

Chistova, S. (2021). Abbreviation in the Russian, English and German discourse of pop music, *Research Result. Theoretical and Applied Linguistics*, 7 (1), 92-115. DOI: 10.18413/2313-8912-2021-7-1-0-8 *(In Russian)*

Cornaggia-Urrigshardt, A., Gökgöz, F., Kurth, F., Schmitz, H. and Wilkinghoff, K. (2022). Speech Recognition Lab, *Procedia Computer Science*, 205, 218–228. https://doi.org/10.1016/j.procs.2022.09.023 (*In English*)

Deng, L. et al. (2013). Recent advances in deep learning for speech research at Microsoft, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8604–8608. DOI: 10.1109/ICASSP.2013.6639345 (*In English*)

Gruzdev, D. and Biktimirov, A. (2022). Written translation via sight translation, *Moscow University Translation Studies Bulletin*, 1, 7-26. (*In Russian*)

Gruzdev, D., Gruzdeva, L. and Makarenko, A. (2019). Sight translation coupled with voice recognition as a key to faster and easier translation, *Bashkir University Bulletin*, 24 (2), 430-438. (*In Russian*)

Jorge, J., Giménez, A., Baquero-Arnal, P., Iranzo-Sánchez, J., Pérez, A., Garcés Díaz-Munío, G.V., Silvestre-Cerdà, J.A., Civera, J., Sanchis, A. and Juan, A. (2021). MLLP-VRAIN Spanish ASR Systems for the Albayzin-RTVE 2020 Speech-To-Text Challenge, *Proceedings of the 5th International Conference "IberSPEECH 2021"*, Valladolid, Spain, 118-122. https://doi.org/10.21437/IberSPEECH.2021-25 *(In English)*

Kol´aˇr, J. and Lamel, L. (2012) Development and evaluation of automatic punctuation for French and English speech-to-text, *Proceedings of the 13th Annual Conference of the International Speech Communication Association "Interspeech 2012"*, Portland, Oregon, USA, 1376-1379. *(In English)*

Kumar, R., Gupta M. and Sapra, S. R. (2021) Speech to text community application using natural language processing, *5th International Conference on Information Systems and Computer Networks (ISCON)*, 1-6. DOI: 10.1109/ISCON52037.2021.9702428 *(In English)*

Kurzekar, P., Deshmukh, R., Waghmare, V. and Shrishrimal, P. (2014). A comparative study of feature extraction techniques for speech recognition system, *International Journal of Innovative Research in Science, Engineering and Technology*, 3 (12), 18006-18016. DOI: 10.15680/IJIRSET.2014.0312034 *(In English)*

Kuzmin, A. and Ivanov, S. (2021). Speech to text system for noisy and quiet speech, *Journal of Physics: Conference Series*, 2096, 012071. https://doi.org/10.1088/1742-6596/2096/1/012071 *(In English)*

Ma, Y., Nguyen, T. H. and Ma, B. (2022). CPT: cross-modal prefix-tuning for speech-to-text translation, *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6217-6221. DOI: 10.1109/ICASSP43922.2022.9746935 *(In English)*

Messaoudi, A., Haddad, H., Fourati, C., Hmida, M. B., Elhaj Mabrouk, A. B. and Graiet, M. (2021). Tunisian Dialectal End-to-end Speech Recognition based on DeepSpeech, *Procedia Computer Science,* 189, 183–190. https://doi.org/10.1016/j.procs.2021.05.082 *(In English)*

Nugraha, D. S. and Dewanti, R. (2022). English-Indonesian crisis translation: accuracy and adequacy of Covid-19 terms translated by three MT tools, *Research Result. Theoretical and Applied Linguistics*, 8 (1), 122-134. https://doi.org/10.18413/2313-8912-2022-8-1-0-8 *(In English)*

Ogunshile, E., Phung, K. and Ramachandran, Raj. (2021). Exploring a web-based application to convert Tamil and Vietnamese speech to text without the effect of code-switching and code-mixing, *Programming and Computer Software,* 47, 757–764. https://doi.org/10.1134/S036176882108020X *(In English)*

Perero-Codosero, J. M., Espinoza-Cuadros, F. M. and Hernández-Gómez, L. A. (2022). A comparison of hybrid and end-to-end ASR systems for the IberSpeech-RTVE 2020 speech-to-text transcription challenge, *Applied Sciences*, 12 (2), 903. https://doi.org/10.3390/app12020903 *(In English)*

Pernarčić, M. (2019). Testing the efficiency of voice recognition software in translation, Master's thesis, Strossmayer University, Croatia. *(In English)*

Stubna, P. (2020). Beyond «Listen and Repeat»: Investigating English Pronunciation Instruction at the Upper Secondary School Level in Slovakia by R. Metruk: A Book Review, *Journal of Language and Education*, 6 (4), 216-220. https://doi.org/10.17323/jle.2020.10919 (*In English*)

Trabelsi, A., Warichet, S., Aajaoun, Y. and Soussilane, S. (2022). Evaluation of the efficiency of state-of-the-art Speech Recognition engines, *Procedia Computer Science*, 207, 2242-2252. https://doi.org/10.1016/j.procs.2022.09.534 (*In English*)

Appendix

Excerpts with the greatest number of errors

Original	Mic	Smartphone	Recorder
A federal judge has found that a former Afghan militant has been held unlawfully at the Guantánamo Bay detention camp, the first time in 10 years that a detainee has won such a case against the U.S. government, his lawyers said. U.S. District Judge Amit P. Mehta in Washington on Tuesday entered a final order and two classified opinions on Asadullah Haroon Gul's petition for a writ of habeas corpus and immediate release, court	A federal judge has found that **(a)** former Afghan militant has been held unlawfully at **(the)** Guantánamo Bay detention camp the first time in 10 years that **ADD** has **one** such a case against the US government his lawyers said US district judge ***Annie eat nectar*** in Washington on Tuesday entered ***the*** final ***oral*** and two classif**y** opinions on ***a similar dance*** petition for a writ of habeas **(cor)*opus*** and immediate release ***caught*** filings	**(A)** federal judge has found that a former Afghan militant has been held unlawfully at the Guantánamo Bay detention camp. The first time in 10 years that **(a)** detainee has ***one*** such **(a)** case against the US government. His lawyers said US District Judge ***Annie Pat met*** in Washington on Tuesday entered ***the*** final order and ***to*** classified opinions on ***a similar Korean gals*** petition for a writ of habeas corpus and immediate release court	a federal judge has found that **(a)** former Afghan militant has been held unlawfully at the Guantánamo Bay detention camp, the first time in 10 years that **(a)** detainee has won such a case against the US government. ***She is***, lawyers said US district judge. ***I need be nectar*** in Washington on Tuesday entered **the** final order and **to** classify**(ied)** **convenience** on ***this underlying her own gal's*** petition for a writ of habeas corpus and

filings confirmed, without disclosing their contents.	confirmed without disclosing the**ir** contents	filings confirmed without disclosing their contents.	immediate release court filings confirmed without disclosing the contents
The control tower is seen through the razor wire inside the Camp VI detention facility in Guantánamo Bay Naval Base, Cuba. (Alex Brandon/AP) Gul's counsel Mark Maher, with the nonprofit group Reprieve, said the lawyers were thrilled for their client.	the control tower is seen through the razor the razor wire inside the camp six Detention facility in Guantanamo Bay__naval base Cuba Alex ***Brendan*** /AP ***Gulf Council*** Mark ***Meyer*** **(with)** the nonprofit group ***green*** said **(the)** lawyers were thrilled for their client	the control tower is seen through the razor wire inside the camp six detention facility in Guantánamo Bay naval base **(in)** Cuba, Alex Brandon/AP ***gal's*** counsel Mark ***Meyer*** with **a** non-profit group reprieve, said the lawyers were thrilled for their client,	the control tower is seen through the razor wire inside the camp VI detention facility in Guantánamo Bay naval base, Cuba, Alex ***Brendan***/AP ***Gulf*** Council, Mark ***Meyer***, with the nonprofit group reprieve, said the lawyers were thrilled for their client,

Конфликты интересов: у автора нет конфликта интересов для декларации.

Conflicts of Interest: the authors have no conflict of interest to declare.

Andrey R. Biktimirov, Ph.D. Student at the English Department, Military University, Russia.

Dmitry Yu. Gruzdev, Ph.D. in Linguistics, Associate Professor, Deputy Head of the English Department, Military University, Russia.