DOI https://doi.org/10.15407/csc.2023.04.019
UDC 004.934

M.M. SAZHOK, Ph.D. (Eng.), head of the department, International Research and Training Center for Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0000-0003-1169-6851, sazhok@gmail.com

V.V. ROBEIKO, Research fellow, Taras Shevchenko National University of Kyiv, Glushkov ave., 4g, 03022, Kyiv, Ukraine, ORCID: https://orcid.org/0000-0003-2266-7650, valya.robeiko@gmail.com

YE.A. SMOLIAKOV, International Research and Training Center for Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0000-0002-8272-2095, egorsmkv@gmail.com

T.O. ZABOLOTKO, International Research and Training Center for Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0009-0002-1575-3091, wariushas@gmail.com

R.A. SELIUKH, Research Associate, International Research and Training Center for Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0000-0003-2230-8746, vxm112@gmail.com

D.YA FEDORYN, Research Associate, International Research and Training Center for Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0000-0002-4924-225X, dmytro.fedoryn@gmail.com

O.A YUKHYMENKO, Research fellow, International Research and Training Center for Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine, ORCID: https://orcid.org/0000-0001-5868-8547, enomaj@gmail.com

MODELING DOMAIN OPENNESS IN SPEECH INFORMATION TECHNOLOGIES

The paper examines the topical problem of applying speech-to-text systems for different domains, including a variety of acoustic conditions, individual characteristics and context types, as well as taking into account multi- and cross-lingual properties. The described techniques for modeling speech signal deterioration, lexicon flexibility and code-switch made it possible to increase the robustness of previously developed systems, which expanded the domains of their application.

Keywords*: speech signal analysis, speech recognition, multilingual processing, speech understanding, domain broadening, punctuation restoration, text decoration, speaking failures.*

Introduction

When developing speech technologies, a number of human intellectual functions are modeled, which provide:

- automatic perception of information transmitted by voice through various communication channels in many languages;
- the ability to search for information in the broadcast stream by text request;
- receiving an answer to the question "Who is speaking?";
- automation of documentation of media sources of information;
- release of hands when entering information (dictation);
- release of the visual channel (voicing of messages and other arbitrary text).

In previous works [7, 11, 13] the authors have developed speech signal transcription systems within subject areas in conditions of user cooperation with the system (text input under dictation), as well as in conditions when individuals do not expect that their speech will be automatically converted into text and communicate spontaneously (transcribing records of meetings or news stories).

Automatic transcription of audio recordings obtained in conditions of notable noise and disturbances allows to document episodes and determine:

- what is said (speech-to-text conversion);
- who is speaking (voice identification and identification of unknown persons);
- what language and cultural context is used (language, country, speaker social position);
- what is the content (classification by topic, sentiment, specific entity detection, etc.);
- what other kinds of sounds are observed and might be interpreted.

The resulting transcription is an array of text synchronized with audio (and video) stored on the server and searchable by text query.

A number of the limitations of the listed works are related to the focus on specific subject areas. Among the shortcomings are: (1) significant sensitivity to acoustic conditions distinguishing from the conditions in the training sample, (2) the inability to fully process proper names and specific vocabulary terms and (3) lack of real multilingualism.

Therefore, the next step of the research is to strengthen independency on the subject area and to model robustly the following:

- acoustic tract properties through which the speech signal is transmitted,
- the (virtual) openness of the lexicon, including speaking failures,
- language change (code-switching).

The next section overviews the recent evolution of approaches used in speech signal analysis and the state-of-art techniques and implementations, the Research and Development section describes applied techniques and developed technology and systems, main assessments are presented and discussed in the Result section, finally, we conclude and spot future research.

Related work

Previous decade was characterized by the dominance of architectures based on Hidden Markov Models (HMM) combined with Gaussian Mixture Models (GMM) with the gradual displacement of GMM by Deep Learning techniques. The typical stages involved in speech-to-text conversion include:

- *Feature Extraction*: The raw audio signal is transformed into a feature representation that captures relevant information for speech recognition. Commonly used features include Mel-Frequency Cepstral Coefficients (MFCCs), filter banks, or spectrograms.
- *Acoustic Modeling*: Acoustic modeling aims to capture the relationship between the acoustic features extracted from the audio signal and the corresponding phonemes or other chosen basic sub-word units. HMMs are commonly used in this stage to model the temporal distortion of basic speech units and GMM is used to model the approximation in feature space.
- *Pronunciation Modeling*: Pronunciation modeling deals with the mapping between phonemes and word-level tokens. This stage typically involves a pronunciation dictionary that maps word-level

tokens to sequences of phonemes, along with rules for handling variations in pronunciation.

▪ *Language Modeling*: Language modeling focuses on predicting the likelihood of word sequences in a given language. This is typically done using statistical language models, such as n-gram models or more sophisticated techniques like recurrent neural networks (RNNs).

▪ *Decoding*: Decoding involves searching through the space of hypothetical word sequences to find the most likely sequence given the acoustic, pronunciation and language models integrated in Weighted Finite State Transducers (WFST). Algorithms based on Dynamic Programming, such as Viterbi decoding or beam search, are commonly used for this purpose.

Nowadays, Deep Learning techniques mostly absorb Acoustic, Pronunciation and Language Modeling stages and are presented by networks composed of multiple layers of interconnected artificial neurons, which can model complex patterns and grasp wide contexts in data.

▪ *Recurrent Neural Networks* (RNNs): RNNs are a type of neural network particularly well-suited for sequential data, such as speech. They have connections that loop backward, allowing information to persist. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, which are variants of RNNs, are often used to handle long-range dependencies and alleviate the vanishing gradient problem.

▪ Convolutional Neural Networks (CNNs): CNNs are highly effective for processing grid-like data, such as images. In speech recognition, they are often used for feature extraction from spectrograms or other time-frequency representations of audio signals. CNNs can capture local patterns efficiently.

▪ *Attention Mechanisms*: Attention mechanisms have been incorporated into speech recognition models to help focus on relevant parts of the input during decoding. They enable the model to align input features with output labels dynamically, improving performance, especially in noisy or complex audio environments. One of the most powerful architectures based on Attention Mechanism is Transformer.

▪ *Connectionist Temporal Classification* (CTC): CTC is a method used for training neural networks to optimize sequence-to-sequence mappings. It's particularly useful in tasks where the alignment between the input and output sequences is not known beforehand, which is common in speech recognition.

▪ *Transfer Learning and Pre-training*: Transfer learning techniques, where models are pre-trained on large datasets (such as LibriSpeech or GigaSpeech), have been instrumental in improving the performance of speech recognition systems, especially in low-resource scenarios.

Normally, the Decoding stage operates with subword token observation scores at each time frame and produces the final sequence of word-level tokens, which may include punctuation as well as letter capitalisation and digital and symbolic presentation of word subsequences. Since the granularity of time frames has become larger, the beam search is used mainly to compose the most probable word-level tokens rather than to approximate nonlinear temporal distortions of speech signal segments.

The observed tendency to combine the traditional stages into a single neural network architecture leads to End-to-end models simplifying the pipeline and potentially improving performance.

The most productive toolkits that implement speech analysis techniques include Kaldi, WeNet, ESPNet, Wav2Vec, Nemo, Whisper, Seamless [4, 6, 8, 10]. Normally, a toolkit includes baseline models derived from train samples based on either publicly available corpora like LibriSpeech, Tedlium, GigaSpeech, VoxCeleb, Multi_CN etc. or some closed corpora. A choice to train a model from scratch or fine-tune a pre-trained model gives a chance to build a viable model with less annotated training data and rather modest computational resources.

Due to the diversity of approaches implemented in the toolkits and train samples annotation the resulting text may include some particular features:

▪ letter capitalisation, punctuation, digits and symbols,

▪ elements of distinct text reconstruction with alleviating of speaking failures,

▪ multilinguality elements.

In the case of word sequence output, the further application of certain techniques may help to restore the required properties from the raw text like it is done in [12, 13].

Speaking failures might be falsely detected at actually correctly spoken segments, which can be crucial. E.g., when the speaker intentionally spells the same digit several times the result may contain only one sample of the digit, having processed the rest of samples as a restart or hesitation.

Multilingual features are available for systems trained on multilingual data like provided in Whisper or Seamless. However, this does not mean that a single inference will produce the result in a code-switch manner. In fact, the text output inference is based on the only selected language, and words in segments containing other languages are, hypothetically, either interpreted or approximated with similarly sounding words of the selected language. Moreover, in certain cases, an unacceptable text output might be produced by end-to-end architectures, which even got a specific term of *hallucinations*. Therefore, to provide real multilingual features we should envisage a code-switch mechanism that should show acceptable results for predefined sets of languages. Currently, bilingual Automatic Speech Recognition (ASR) systems can be found only in commercial production for such language pairs as English–Spanish and English–Chinese [1 3].

Noisy, distorted and overlapped speech causes the significant deterioration of recognition results. A huge progress in speech enhancement techniques is promising for recognition improvement in noisy and distorted speech segments as well as in segments with overlapped speech [9, 15].

Research and development

In this work the numerous techniques and their implementations were analyzed, tested and combined into pipelines of means providing the functionality of speech recognition systems aimed to alleviate a substantial domain dependency.

The properties of the acoustic tract are determined by various causes like external noises, reverberation, interference, simultaneous speech of several people, speaking styles and the state of the speaker (emotion, singing), lossy compression etc. This is topical for domains related to telephony and wireless analogous communication means as well as for records done outdoors by a highly sensitive microphone.

Modeling of the acoustic tract takes place on the basis of a training set that has one or more of the listed properties, or due to augmentation – artificial provision of the initial signal of the training sample with specified properties [16].

The most challenging among the listed issues is modeling the overlapping of signals of two and more speech sources [9, 15]. It is even not obvious how to transcribe overlapped segments. This problem was not paid proper attention due to its relatively infrequent occurrence until practically absence for certain specific tasks like phone calls transcribing with isolated records for each call participant. However, when the same records are mixed into a single channel, nothing prevents overlapped speech segments from occurring.

Lexical limitation that ties to the domain cannot be avoided in speech-to-text systems with a closed dictionary, no matter how large the dictionary is. For example, for Ukrainian language, non-dictionary word forms make up an average of 3–5% for a dictionary volume of 200,000 words. Moving to dictionaries of sub-word segments like morphemes instead of words makes it possible to alleviate restrictions on the lexicon significantly providing its kind of virtualization.

Considering morphemes as a basic unit instead of phonemes leads us to direct estimations of observation probabilities for parts of words by the multilayer network. In Fig. 1 a sample of Ukrainian word “oca” can be segmented by recognized subword hypotheses that include: "o", "oc", "c", "ca", each of which has its probability-based score at time frames. While decoding word sequences, an optimal or suboptimal path is computed for all permissible sequences of subword units by means of the beam search with optional involving certain language model scores.

Another application of subwords is handling of speaking failures like false-starts, self-corrections and hesitations. Each hypothetical word can be

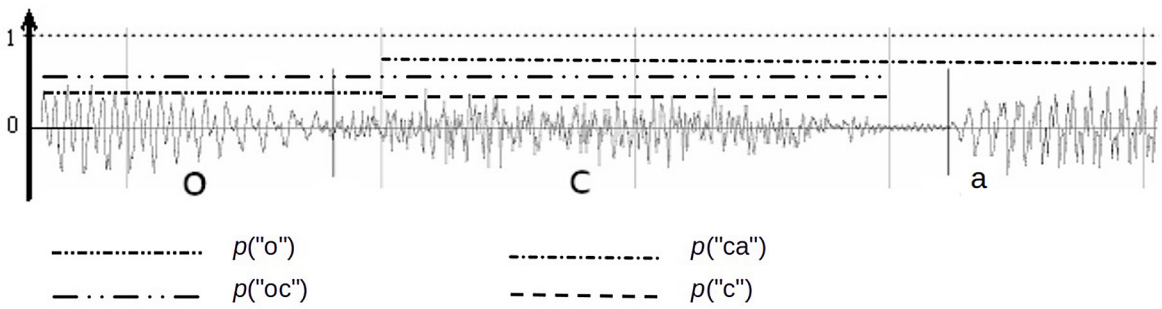

Fig. 1. Subword segmentation with token score visualization

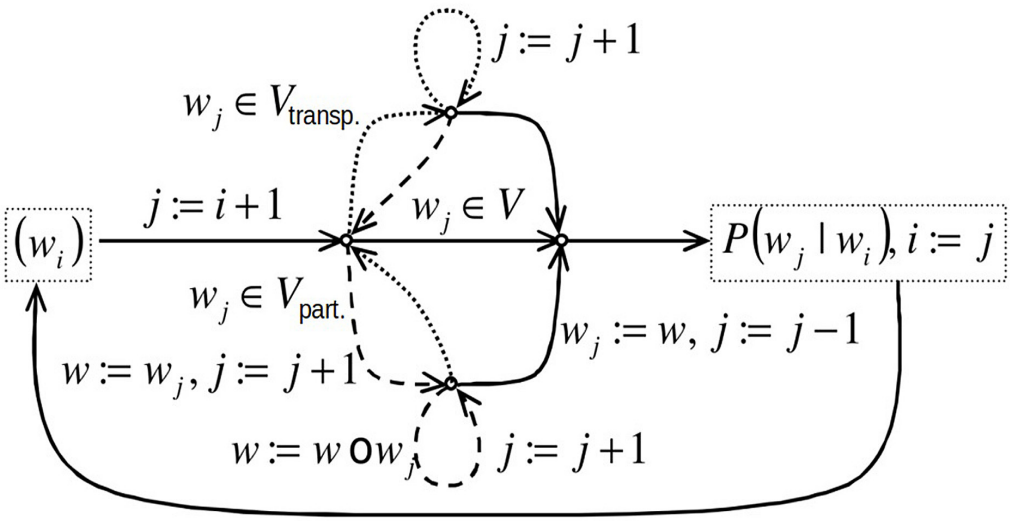

Fig. 2. Modeling of speech failures

divided into partial words and one or more filled pauses, which are invisible (transparent) in the decoded word sequence. So, in addition to the virtual dictionary of words, V, we introduce dictionaries of all partial words, V_{part}, and filled pauses, V_{transp}. Then, upon arrival of a hypothetical partial word, we can delay the accumulation of the full word until the completion of following the hypothetical filled pauses. Fig. 2 shows the graph by which a bigram of $P(wj \mid wi)$ is formed for its further score adjustment. Dashed lines indicate the accumulation of a partial word, the dotted line corresponds to the arrival of a filled pause, which is dropped. A circle operator means a concatenation of subword segments. In the general case, it is possible to advance with two arrows at the same time, when the full word coincides with the partial word w: $w \in V \cap V_{part} \neq \varnothing$.

Code-switching means alternation of different languages between or within sentences for different or the same speaker. In countries where more than one language is spoken by a significant number of people the code-switching speech is a common phenomenon. In Ukraine, switches, as well as language mixtures (surzhyk), are often observed in bilingual spontaneous speech. Solving the problem is carried out in two directions: (1) creation of a common dictionary with combined sets of basic speech units or (2) parallel recognition of signal segments

Fig. 3. Automatic speech transcribing basic web-interface

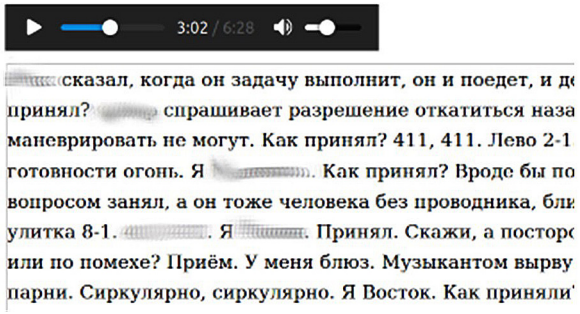

Fig. 4. Speech-to-text conversion result for a set of intercepted radio transmissions

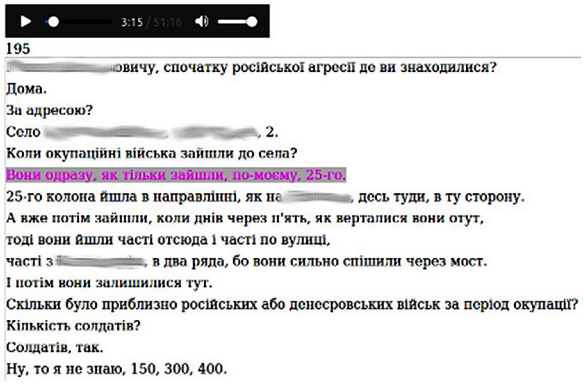

Fig. 5. Speech-to-text conversion for investigative action documenting

using monolingual models with subsequent merging of separate results into one [1, 2].

Speech-to-text conversion system output consists of words composed of basic tokens. When we consider tokens including only orthographic case-independent presentation, then we need to introduce one or more text decorating models that cover word-to-number conversion as well as punctuation and case restoration. Number extraction is provided by word sequences consisting of certain basic numeric words. The authors use a rule-based sequence-to-sequence conversion procedure, similar to grapheme-to-phoneme converter, with introduced multilingual rules allowing for word sequence segmenting into numeric and generic words and extracting digital spelling for any integer number [5, 12]. For punctuation restoration, we apply a model based on Recurrent Neural Nets, where the encoder sequentially encodes word sequences optionally accomplished with prosody, and the decoder decodes the text with restored punctuation. The model parameters were estimated on a Ukrainian text corpus containing about 10 million sentences extracted from news web-sites so no prosody features were used [13].

Before restoring punctuation marks and converting word sequences to numbers, a speaker turn changing detection, i.e., speaker diarization, should be accomplished. We assume that when the turn to speak is transferred to another party, the next thought is started, so a new sentence begins. Rare cases, when other conversation parties can continue or finalize the sentence, are not currently considered.

After the speaker diarization stage is passed, we recover the numbers from the word sequences and then restore the punctuation. This is due to the fact that the parameters of the punctuation model are more appropriate for estimation and evaluation on texts containing numeric expressions rather than word sequences. Otherwise, before estimating the parameters, it would be necessary, in the training text sample, to convert the numeric expressions to a word sequence, which is not a sufficiently analyzed and solved problem, particularly for Ukrainian language because of necessity to model the gender and case agreement.

With the use of appropriate open source tools and multilingual speech and text corpora either open, like Common Voice and UberText, or closed, the experimental system have been developed in the client–server architecture with a basic web-interface [4, 8, 10, 12]. Fig. 3 illustrates the input interface where a user can choose a record among local files or provide a link to media data,

adjust desired media segment boundaries and select a language or mixture of languages. The application of this system is intended for the event fixation documentation automating, audiorecord transcribing and subsequent interpretation and analyzing of telecommunication and other media sources (television, radio, youtube, contact centers, etc.) by recognized text.

The example of automatic transcription for a set of intercepted radio transmissions is shown in Fig. 4. This is a highly coarse noise signal with specific lexicon. Here, the speech-to-text conversion result might be used to analyze the current situation operatively due to automating.

Fig. 5 illustrates the speech-to-text conversion of an investigative action that might be used as a basis for documenting the events of legal nature. Since the record is made by a highly sensitive microphone, external noises and speech overlaps are observed. The inter-sentential code-switching with elements of surzhyk can be found in line 9, which is processed as expected.

Experimental results

To assess speech-to-text conversion, firstly, for the developer set, the least error-prone hyper-parameters, in terms of word error rate (WER), are being searched and adjusted, then WER is estimated for the prepared test sets.

The test samples were taken from both publically available and closed sources. The crowd-sourced freely available Ukrainian part of the Common Voice set of corpora is continuously collected by volunteers and, at the experimenting time, contained 12 hours of isolated phrases. Phrases and respective annotations were examined and about 2 hours of records were removed from the sample as garbage data. The bilingual test set was collected from noisy sources, particularly from telephonic channel, and 3 test samples were prepared on its basis:

- mixed inbound and outbound channels,
- inbound and outbound speech placed into different files and
- mixed signal was automatically divided by source and stored to respective files.

Table 1. **Speech-to-text monolingual test accuracy, %WER**

Test set \ Model	Bilingual	Monolingual	Monolingual + LM	Whisper, large v2
Common Voice, Ukrainian, 10h	10,7	7,2	6,5	13,5

Table 2. **Speech-to-text bilingual test accuracy, %WER**

Test set \ Model	Bilingual
Bilingual, telephonic, mixed	17,2
Bilingual, telephonic, stereo	16,1
Reconstructed stereo	16,6

Table 3. **Punctuation restoration results in F1-score, %**

Test set \ Model	Monolingual baseline	Monolingual improved	Bilingual
Monolingual, artificial	68,1	73,5	71,7
Bilingual	61,9	63,5	66,9

The bilingual test samples contain spontaneous conversations with sometimes emotional speech and noises and cover inter- and intra-sentential code-switching. The total duration of multilingual test data is about 5 hours.

Model parameters were estimated in single language (monolingual) and merged languages (bilingual) modes [13]. To estimate parameters for the monolingual model about 1000 hours of speech were taken and almost 2400 hours of speech were used for the bilingual model. Language model (LM) was trained only for monolingual mode using text data extracted mostly from news web-sites with a total amount of 2GB.

A publically available Whisper model taken for the comparison is trained on the multilingual closed corpus containing 680K hours of speech with 1550M parameters estimated [10].

The developed models showed better accuracy for isolated phrases and for the monolingual setup LM application gave a significant improvement (Table 1). The WER below 20% reported in Table 2 for the only bilingual model confirms that the developed model enables effective assisting for speech transcribing as well as a further analysis of the rec-

ognized text by NLP automatic means under the conditions of domain extension.

The WER reported by the Whisper team for the similar test set is slightly different, as expected, which can be explained by the cleaning procedure applied with thorough supervision to the initial test data.

The punctuation restoration model is evaluated in terms of F1-score, which is a geometric mean of precision and recall for such punctuation marks as period, comma and question mark. The monolingual test set consists of 32000 words and is obtained by removing the punctuation marks from text. Bilingual test set is based on real texts making a total of 12000 words. The baseline monolingual model is the result of previous work [7, 13]. Other models were trained on an updated training set: 2 GB for the improved monolingual model and 2.5 GB for the bilingual model.

As shown in Table 3, language merging is beneficial for multilingual punctuation restoration.

Conclusions

The achieved accuracy of the developed model shows weaker sensitivity to acoustic conditions and various contexts, and allows for processing real speech recorded for Ukrainian cultural environment.

The implementation of the described approaches to modeling the domain openness makes it possible to move on to the creation of speech signal transcription systems that convert almost any speech information into text that is compact and convenient for human reading and further NLP processing. Particularly, according to the acquired text, topics and sentiments can be extracted, named entities can be tracked (names, dates, measurements, etc.), punctuation marks improve the perceptual experience of the text and, in general, the time and cost of manual editing to get a record properly documented can be considerably reduced.

Future research includes:

- speaker identification implementation;
- episode segmentation;
- tonality and toxicity detection by both text and speech acoustic cues;
- applying acoustic cues to punctuation restoration;
- recognition of specified noise types (laugh, applause etc.);
- text reconstruction that operates with wider context;
- multimodal speech recognition.

Acknowledgment

This study was partially supported by the National Academy of Science of Ukraine under the research project titled: “Information technology for acoustic and subject environment openness modeling in spontaneous language between a human and cybernetic systems”, State Registration No 0121U000015.

REFERENCES

1. Ugan, E.Y., Huber, C., Hussain, J., Waibel, A. (2023). “Language-agnostic Code-Switching in Sequence-To-Sequence Speech Recognition”. arXiv preprint arXiv:2210.08992v2 [cs.CL], 3 Jul 2023.
2. Lyudovyk, T., Pylypenko, V. (2014). “Code-switching speech recognition for closely related languages. In Spoken Language Technologies for Under-Resourced Languages”. International Research and Training Center for Information Technologies and Systems, Kyiv, Ukraine.
3. Lovenia, H., Cahyawijaya, S., Winata, G. I., Xu, P., Yan, X., Liu, Z., ..., Fung, P. (2021). ASCEND: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. arXiv preprint arXiv:2112.06223. The Hong Kong University of Science and Technology. [cs.CL], 3 May 2022.
4. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ..., Schwarz, P. (2011). “The Kaldi speech recognition toolkit. IEEE 2011 workshop on automatic speech recognition and understanding”. IEEE Signal Processing Society. IEEE Catalog, No.: CFP11SRW-USB.
5. Sazhok, M., Robeiko, V. (2013). “Lexical Stress-Based Morphological Decomposition and Its Application for Ukrainian Speech Recognition”. In Text, Speech, and Dialogue: 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings 16, pp. 327-334. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40585-3_42

6. Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., ..., Ochiai, T. (2018). Espnet: End-to-end speech processing toolkit. Proc. Int. conference Interspeech'2018. arXiv preprint arXiv:1804.00015. https://doi.org/10.21437/Interspeech.2018-1456
7. Sazhok M.M., Seliukh R.A., Fedoryn D.Ya., Yukhymenko O.A., Robeiko V.V. (2019). "Automatic Speech Recognition For Ukrainian Broadcast Media Transcribing". Control Systems and Computers, No6 (264), pp. 46-57. https://doi.org/10.15407/csc.2019.06.046
8. Yao, Z., Wu, D., Wang, X., Zhang, B., Yu, F., Yang, C., ..., Lei, X. (2021). "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit". In Proc. Int conference Interspeech' 2021, Brno, Czechia. arXiv preprint arXiv:2102.01547. https://doi.org/10.21437/Interspeech.2021-1983
9. Lu, Y. J., Chang, X., Li, C., Zhang, W., Cornell, S., Ni, Z., ..., Watanabe, S. (2022). "ESPnet-SE++: Speech enhancement for robust speech recognition, translation, and understanding". In Proc. Int conference Interspeech' 2022, pp. 5458-5462. arXiv preprint arXiv:2207.09514. https://doi.org/10.21437/Interspeech.2022-10727
10. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, pp. 28492-28518. PMLR. arXiv:2212.04356 [eess.AS], 2022, https://arxiv.org/abs/2212.04356
11. Vintsiuk T.K., Sazhok M.M., Seliukh R.A., Fedoryn D.Ya., Yukhymenko O.A., Robeiko V.V. (2018). "Automatic recognition, understanding and synthesis of speech signals in Ukraine". Control Systems and Computers. No 6 (278), pp. 7-24, https://doi.org/10.15407/usim.2018.06.007 (In Ukrainian).
12. Sazhok, M., Seliukh, R., Fedoryn, D., Yukhymenko, O., & Robeiko, V. (2020). "Written form extraction of spoken numeric sequences in speech-to-text conversion for Ukrainian". In CEUR workshop proceedings, pp. 442-451. https://ceur-ws.org/Vol-2604/paper32.pdf
13. Sazhok, M.M., Poltyeva, A., Robeiko, V., Seliukh, R., Fedoryn, D. (2021). "Punctuation Restoration for Ukrainian Broadcast Speech Recognition System based on Bidirectional Recurrent Neural Network and Word Embeddings". In Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems 2021 (COLINS-2021), pp. 300-310. https://ceur-ws.org/Vol-2870/paper25.pdf
14. Zasukha, D. (2023). "Using Thumbnail Length Bounds To Improve Audio Thumbnailing For Beatles Songs". Stucnij intelekt. 2023. 28 (1). pp. 60-65 (In Ukrainian) [Засуха Д. Використання границь довжини мініатюри для покращення процедури отримання звукової мініатюри пісень Бітлз. Штучний інтелект. 2023. 28 (1). pp. 60-65]. https://doi.org/10.15407/jai2023.01.060
15. Manuel Pariente, Samuele Cornell, Joris Cosentino et al.. (2020). Asteroid: the PyTorch-based audio source separation toolkit for researchers. Proc. Int. conference Interspeech'2020. arXiv preprint arXiv:2005.04132. https://doi.org/10.48550/arXiv.2005.0413
16. Mina Huh, Ruchira Ray, Corey Karnei. (2023). A Comparison of Speech Data Augmentation Methods Using S3PRL Toolkit. arXiv preprint arXiv:2303.00510, https://doi.org/10.48550/arXiv.2303.00510

Received 09.11.2023

М.М. Сажок, канд. техн. наук, зав. відділом, Міжнародний науково-навчальний центр
інформаційних технологій та систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ, 03187, Україна,
ORCID: https://orcid.org/0000-0003-1169-6851,
sazhok@gmail.com
В.В. Робейко, науковий співробітник, Київський національний університет
імені Тараса Шевченка, 03022, Київ, просп. Академіка Глушкова, 4,
ORCID: https://orcid.org/0000-0003-2266-7650,
valia.robeiko@gmail.com
Є.А. Смоляков, молодший науковий співробітник, Міжнародний науково-навчальний
центр інформаційних технологій і систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ, 03187, Україна,
ORCID: https://orcid.org/0000-0002-8272-2095,
egorsmkv@gmail.com

Т.О. Заболотько, науковий співробітник, Міжнародний науково-навчальний центр
інформаційних технологій і систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ, 03187, Україна,
ORCID: https://orcid.org/0009-0002-1575-3091,
wariushas@gmail.com
Р.А. Селюх, молодший науковий співробітник, Міжнародний науково-навчальний центр
інформаційних технологій та систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ, 03187, Україна,
ORCID: https://orcid.org/0000-0003-2230-8746,
vxml12@gmail.com
Д.Я. Федорин, молодший науковий співробітник, Міжнародний науково-навчальний
центр інформаційних технологій та систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ, 03187, Україна,
ORCID: https://orcid.org/0000-0002-4924-225X,
dmytro.fedoryn@gmail.com
О.А.Юхименко, науковий співробітник, Міжнародний науково-навчальний центр
інформаційних технологій та систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ, 03187, Україна,
ORCID: https://orcid.org/0000-0001-5868-8547,
enomaj@gmail.com

МОДЕЛЮВАННЯ ВІДКРИТОСТІ ПРЕДМЕТНОЇ ОБЛАСТІ В МОВЛЕННЄВИХ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЯХ

Вступ.
У попередні роки розроблені системи транскрибування мовленнєвого сигналу в межах предметних областей, як в умовах кооперування користувача з системою (введення тексту під диктування), так і в умовах, коли особи не розраховують, що їх мовлення буде автоматично перетворене на текст, і комунікують спонтанно (транскрибування фонограм засідань або сюжетів новин). До недоліків цих розробок варто віднести: значну чутливість до акустичних умов, відмінних від умов у навчальній вибірці, неможливість опрацьовувати повною мірою власні назви, особливостей лексики та вимови та недостатність моделювання властивої Україні багатомовності. Постало завдання відійти від прив'язаності до предметної області, що головним чином характеризується: акустичним трактом, через які передається мовленнєвий сигнал, присутністю в лексиці специфічних слів, розмаїттям індивідуальних особливостей і міжмовним перемиканням.

Мета. Розробити технологію, що дає змогу ефективно розпізнавати мовленнєвий сигнал в розмаїтому предметному і акустичному середовищі, знявши якомога більше обмежень, зумовлених необхідним раніше адаптуванням до конкретної предметної області.

Методи. Моделювання розмаїтості акустичного тракту відбувається на підставі навчальної вибірки, що має одну і більше перелічених властивостей, або за рахунок огментації – штучного надання початковому сигналу навчальної вибірки заданих властивостей. Перехід до словників із субслівних сегментів-морфів замість слів дало змогу зняти строгі обмеження на лексикон, компонувати віртуально будь-які слова, обробляти мовленнєві збої та використовувати морфи в якості базової одиниці мовлення замість фонем. Вирішення проблеми міжмовного перемикання здійснюється за двома напрямками: (1) створення спільного словника з, відповідно, об'єднаними множинами базових одиниць мовлення або (2) паралельного розпізнавання сегментів сигналу з використанням одномовних моделей з подальшим злиттям окремих результатів в один.

Результати та висновки. З використанням відповідних інструментальних засобів та багатомовних мовленнєвих і текстових корпусів розроблено експериментальні системи в архітектурі клієнт-сервер, які знаходять застосування при автоматизації документування фіксацій подій, транскрибуванні та подальшій інтерпретації і моніторингу телекомунікаційних та медійних джерел (телерадіоефір, youtube тощо). Реалізація підходів до моделювання відкритості предметної області дала змогу перейти до створення таких систем транскрибування мовленнєвого сигналу, що перетворюють на зрозумілий людині текст практично будь-яку мовленнєву інформацію, досягають кращої надійності за меншої чутливості до акустичних умов і різноманітних контекстів, обробляють мовленнєві збої та моделюють реальну багатомовність.

Ключові слова: *мовлення, мова, мовленнєвий сигнал, аналіз, розпізнавання, розуміння, акустичні шуми і завади, відкритість предметної області, багатомовність, мовленнєві збої.*