IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

A Systematic Survey of Multilingual Speech Transcription and Translation

Vaibhav Ravindra, Dheeraj K N, Prof. Vinay Raj
Department of Information Science and Engineering
Global Academy of Technology, Bangalore, Karnataka, India

Abstract: *This paper presents an innovative initiative dedicated to revolutionizing multilingual communication by leveraging state-of-the-art technologies such as Artificial Intelligence (AI), NLP, or natural language processing and Machine Learning. With a primary focus on Indian languages, the research aims to develop an advanced system capable of seamlessly transcribing speech across diverse linguistic landscapes. Through the incorporation of cutting-edge algorithms and parallel processing techniques, the proposed system facilitates real-time transcription and translation of multiple languages concurrently. Rigorous experimentation and analysis demonstrate the efficiency of the developed framework in breaking down language barriers and fostering inclusive communication. Furthermore, the paper emphasizes the cultural significance of this technology in promoting global connectivity and celebrating linguistic diversity. Ultimately, this research underscores the transformative potential of technology in facilitating cross-cultural understanding and enabling meaningful interactions within a multilingual society*

Keywords: Natural Language Processing or NLP, Speech-to-text translation, Linguistic analysis, Unsupervised learning, Hidden Markov Models (HMM), Speech-to-speech translation**.**

I. INTRODUCTION

In recent years,the domain of multilingual speech transcription and translation has witnessed significant advancements and gained increased attention because of its potential to revolutionize language processing and facilitate effective communication across diverse linguistic communities. The ability to translate audible words into written language and seamlessly translate it across different languages has significant consequences for a number of domains, including communication, accessibility, education, and cross-cultural understanding. This survey paper aims to offer a thorough and detailed summary of the body of work already written in the area of multilingual speech transcription and translation systems. By examining a variety of research papers, studies, and advancements, we aim to identify the key ideas, methodologies, challenges, and gaps in the current state-of-the-art approaches. The literature survey encompasses a diverse range of topics and methodologies, offering insights into the latest study patterns and advancements. One prominent area of study is unsupervised speech-to-text translation, which leverages monolingual voice and text corpora to provide translation without requiring tagged data. These methods make use of strategies like bilingual cross-modal dictionaries, autoencoders, and language models to enhance translation quality. Another significant research direction focuses on the implementation of speech-to-text conversion using Hidden Markov Models (HMM). These models aim to enhance text comprehension and provide substantial benefits, particularly for visually impaired users. By leveraging the statistical properties of speech, HMM-based approaches offer promising results in accurately writing down spoken language using transcription.

Furthermore, the survey explores the advancements in end-to-end speech-to-text translation with two-pass decoding. This research direction addresses the intricate interactions between audio in the source language and text in the target language by proposing an end-to-end architecture for speech-to-text translation. The utilization of two decoders has demonstrated improved translation results. Enhancing speech-to-speech translation with multiple text-to-speech (TTS) targets is another area of interest. These approaches utilize a multi-task framework to optimize multiple targets simultaneously, leading to improved translation performance. By studying the impact of various synthetic target speech on voice-to-speech translation models, researchers aim to achieve more accurate and natural-sounding translations. The

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

incorporation of speech-to-text (STT) and text-to-speech (TTS) technologies for educational purposes is another significant research focus. These studies aim to create an elementary emulator that utilizes STT and TTS technologies to develop documentation for English language learners. By examining development trends, application results, and technical overviews, researchers seek to improve the learning experience and accessibility for students.

Moreover, the survey investigates the application of meta-learning in developing modality-neutral multi-task speech translation models. By using meta-learning techniques, these models produce state-of-the-art outcomes for various language pairings, outperforming earlier transfer learning techniques. The emphasis is on achieving high-quality translations while considering ethical considerations and inclusivity. In addition to the above, the survey covers various other aspects of the field, including robust natural language processing, student feedback analysis using NLP, anonymizing speech for speaker privacy, voice-to-text transcription for healthcare organizations, deep learning-based speech recognition and synthesis, parts of speech tagging for different languages, low-resource speech-to-speech translation, and named entity recognition using advanced models like BERT. By systematically reviewing the literature, this survey paper aims to identify the gaps and limitations in the current research landscape. The findings will serve as a foundation for proposing novel approaches and addressing the identified gaps. The proposed work will focus on integrating multilingual transcription and translation into a unified framework without requiring tagged data, improving visibility of users, and exploring user-centric customization and cultural adaptability in translation models. Overall, this survey paper serves as an extensive and an invaluable tool for scholars, professionals, and enthusiasts interested in gaining a comprehensive understanding of the current state-of-the-art, challenges, and future directions in the field of multilingual speech transcription and translation systems. The insights provided can inspire further research and development of innovative approaches, driving the advancement of language processing technologies and promoting effective communication across linguistic barriers.

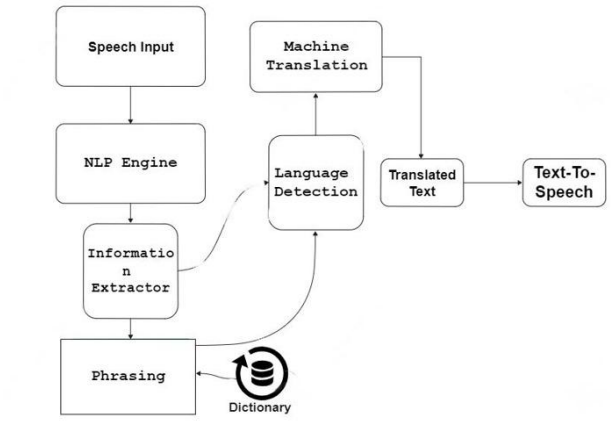

Fig 1.1 System Architecture

II. LITERATURE SURVEY

In [1], a framework is presented for developing speech-to-text translation (ST) systems using only monolingual speech and text corpora. The system initialization involves a cross-modal bilingual dictionary derived from monolingual corpora, enabling word-by-word translation for unseen speech utterances. Experimental results indicate similar BLEU scores to supervised models, rendering it applicable for language duos that have limited resources.

In [2], the authors suggest the utilization of Hidden Markov Models (HMM) for speech-to-text conversion, aiming to improve text understanding and supporting visually impaired users. The synthesized speech aims to deliver a comprehensible output from audio inputs, employing Digital Signal Processing (DSP) algorithms.

In [3], a fresh end-to-end ST framework with two decoders is presented to address deeper connections between source language audio and target language text. By utilizing paired source language audio and target language text in training, The suggested methodology exhibits enhanced performance compared to traditional cascaded systems.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

In [4], the emphasis is on improving speech-to-speech translation (S2ST) models by examining the effect of synthesized target speech. A multi-task framework is suggested to enhance S2ST systems with multiple targets from various text-to-speech (TTS) systems, yielding consistent enhancements over baselines.

In [5] investigates the application of Text-to-Speech (TTS) and Speech-to-Text (STT) technologies in developing a tool for educational documentation in English. The article offers an overview of these technologies, their applications, and development trends, highlighting their significance for language learning.

In [6] presents a meta-learning approach for end-to-end speech translation, aiming to address data scarcity challenges. The suggested approach does this by transferring information from source tasks (ASR+MT) to the target task (ST). cutting-edge results for English German and English French language pairs.

In [7] provides an in-depth evaluation of robustness in natural language processing, discussing various aspects and strategies to fortify NLP systems against adversarial attacks. The article offers insights and suggestions for future research in this domain.

In [8] compiles NLP approaches employed for analysing student feedback to instructors, presenting a methodical review of techniques and trends in this area. The study seeks to aid researchers in organizing their concepts and pinpointing areas for further advancement.

In [9] suggests an approach for anonymizing speech recordings using generative adversarial networks to safeguard speaker privacy while maintaining content intelligibility. The method surpasses prior techniques in privacy and utility, presenting a hopeful solution for privacy-conscious applications.

In [10], a Voice-to-Text transcription system utilizing CMU Sphinx is presented for healthcare organizations, allowing counsellors and NGOs to document conversations during surveys and transcribe them into text. The offline system facilitates multi-language recognition, assisting in data storage and retrieval.

In [11] delivers an extensive overview of Speech-to-Text (STT) and Text-to-Speech (TTS) recognition technologies, showcasing advancements, applications, and challenges. The paper explores the shift to deep learning methods and their influence on communication and user experience.

In [12] tackles Parts of Speech (POS) tagging for Kannada and Hindi languages employing Machine Learning (ML) as well as Deep Learning (DL) models. The research progresses linguistic studies by scrutinizing experiments on a vast corpus and addressing morphological complexities.

In [13] concentrates on low-resource speech-to-speech translation of English videos to Kannada with lip synchronization, endeavouring to narrow language disparities in instructional materials. The proposed system employs ASR, NMT, and TTS algorithms to realize high-quality translations.

In [14] suggests a concatenative approach for Kannada speech synthesis utilizing syllables as fundamental units, capitalizing on the syllable-centric structure of Indian languages to produce high-quality synthesized speech. The research introduces a technique for text analysis, syllabication, and concatenation.

In [15] advocates a BERT-based method for identifying named entities (NER) in Kannada, facilitating diverse applications in information extraction and comprehension. The paper presents an efficient technique for recognizing and categorizing identified entities within an unorganized text.

In [16], the emphasis is on cross-lingual summarization from English to Kannada, introducing a technique termed "Late Translation" to combine summarization and translation. The paper tackles the scarcity of high-quality multilingual resources and presents a technique for improving information accessibility between languages.

In [17] introduces a combined system that uses the advantages of both ASR and LID modules to provide multilingual voice recognition and language identification. The suggested approach attains precise language detection and performance akin to single-language ASR systems.

In [18] examines the utilization of advanced NLP for high-quality text-to-speech synthesis in Bengali, addressing the integration of CNNs into the speech-to-text framework. The research aims to reduce data requirements and enhance the effectiveness of speech synthesis systems.

In [19] explores cross-cultural Computer-aided translation and speech-to-text recognition enable learning activities systems. The research illustrates the viability and efficacy of employing these systems to facilitate dialogue and information sharing between participants from diverse cultures.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

In [20] deals with the acknowledgement of handwritten Kannada words utilizing various Machine Learning (ML) models, concentrating on feature extraction techniques and preprocessing methods. The research achieves high accuracy in recognizing handwritten words and converts them into speech using the gTTS API

III. ANALYSIS TABLE

Sl. No	Paper Title	Key Ideas in earlier research	Gaps in Lit Survey addressed in our endeavors
1	Towards Unsupervised Speech-to-text Translation	Unsupervised method leveraging monolingual data for speech-to-text translation with a bilingual dictionary.	Integrating multilingual transcription and translation into a single system, offering real-time translation without the need for tagged data.
2	Implementation of Speech to Text Conversion Using Hidden Markov Model	UtilizingHidden Markov Models (HMM) for speech-to-text synthesis to benefit visually impaired users.	Incorporating multilingual transcription and translation, enhancing accessibility for users with diverse linguistic backgrounds.
3	Towards end-to-end speech-to-text translation with two-pass decoding	Proposes an end-to-end architecture for speech-to-text translation with improved results using two decoders.	Combining simultaneous multilingual transcription and translation with user-centric customization, providing a more adaptable and accurate translation experience.
4	Enhancing Speech-To-Speech Translation with Multiple TTS Targets	Introduces a multi-task framework for speech-to-speech translation, optimizing multiple targets simultaneously.	Emphasizing cultural adaptability and linguistic diversity, ensuring accurate translation across diverse language pairs and accents.
5	An Elementary Emulator Based on Speech-To-Text and Text-to-Speech Technologies for Educational Purposes	Examines STT and TTS technologies for educational purposes and offers a method for English language learning documentation.	Investigating the integration for recognition of speech and translation into educational feedback systems, facilitating communication and comprehension between students and instructors.
6	End-end speech-to-text translation with modality agnostic meta-learning	Uses meta-learning to create a modality-neutral multi-task speech translation model.	Exploring ethical considerations and inclusivity in speech recognition, addressing privacy concerns and promoting fairness and transparency.
7	Robust natural language processing: Recent advances, challenges, and future directions	Comprehensive assessment of robustness in NLP and recommendations for further research.	Emphasizing user-centric design and personalization of voice recognition and translation systems, enhancing accessibility and usability for diverse users.
8	Natural Language Processing of Student's Feedback to Instructors: A Systematic Review	Synthesizes NLP approaches used in student feedback analysis and identifies areas for further research.	Investigating the personalization of voice recognition and and translation into educational feedback systems, facilitating communication and comprehension between students and instructors.
9	Anonymizing Speech with Generative Adversarial	Proposes a method for anonymizing audio	Exploring the application of speech recognition and translation in privacy-

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

	Networks to Preserve Speaker Privacy	recordings using generative adversarial networks.	preserving technologies, ensuring speaker privacy while maintaining utility and intelligibility.
10	Voice to Text transcription using CMU Sphinx A mobile application for healthcare organization	Proposes an offline voice-to-text transcription solution for healthcare organizations.	Integrating multilingual transcription and translation into a mobile application, providing real-time translation for healthcare professionals and patients.
11	Speech-to-Text and Text-to-Speech Recognition Using Deep Learning	Discusses methods, applications, and challenges of speech-to-text and text-to-speech technology.	Exploring the integration of deep learning techniques in voice recognition and translation, enhancing accuracy and performance.
12	Parts of Speech Tagging for Kannada and Hindi Languages using ML and DL models	Addresses POS tagging for Hindi and Kannada using ML and DL algorithms.	Advancing linguistic study by integrating multilingual transcription and translation with POS tagging, providing deeper linguistic analysis and understanding.
13	Low Resource Speech-to-Speech Translation of English videos to Kannada with Lip-Synchronization	Develops a system for speech-to-speech translation with lip synchronization for English to Kannada videos.	Integrating multilingual transcription and translation with lip synchronization, enhancing the naturalness and usefulness of translated content.
14	Syllable as the basic unit for Kannada speech synthesis	Proposes a concatenative approach for the synthesis of Kannada speech using syllables as basic units.	Exploring innovative approaches to speech synthesis by incorporating linguistic analysis and syllable-centric structures into the translation process.
15	Named Entity Recognition Using BERT Model for Kannada Language	Suggests a BERT-named entity recognition based on technique for Kannada.	Incorporating multilingual transcription and translation with recognized entity identification, facilitating information extraction and comprehension in diverse linguistic contexts.
16	Natural Language Processing based Cross Lingual Summarization	Addresses cross-lingual summarization from English to Kannada using a "Late Translation" technique.	Exploring the application of multilingual transcription and translation in cross-lingual summarization, enhancing information accessibility between languages.
17	A unified system for multilingual speech recognition and language identification	Presents a bilingual LID and ASR system for voice recognition tasks.	Integrating multilingual transcription and translation with dynamic identification of language, improving accuracy and performance in voice recognition tasks.
18	An Implementation of Advanced NLP for High-Quality Text-To-Speech Synthesis	Explores the application of CNN for Bengali text-to-speech synthesis using NLP.	Investigating the integration of deep learning techniques in text-to-speech synthesis, enhancing the quality and naturalness of synthesized speech.
19	Facilitating cross-cultural understanding with learning activities supported by speech-to-text recognition and computer-aided translation	Demonstrates how translation and speech-to-text recognition aid in cross-cultural learning.	Exploring the application of multilingual transcription and translation in cross-cultural learning, promoting communication and understanding between diverse cultural groups.
20	Text to Speech Conversion of	employs models of machine	Integrating multilingual transcription and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

	Handwritten Kannada Words Using Various Machine Learning Models	learning to handwritten Kannada word recognition and text-to-speech conversion.	translation Using models of machine learning, facilitating accessibility and comprehension for Kannada speakers.

IV. CONCLUSION

This thorough review of the literature offers a holistic overview of the advancements and challenges in the field of multilingual speech transcription and translation systems. By examining a broad variety of research papers, the survey highlights the progress made in unsupervised speech-to-text translation, speech-related hidden Markov models synthesis, and end-to-end architectures for improved translation accuracy. The survey also emphasizes the importance of incorporating multiple TTS targets to enhance speech-to-speech translation models, promoting cultural adaptability and accommodating linguistic diversity. It sheds light on the potential applications of speech-to-text and text-to-speech technologies in education, facilitating documentation for language learners and improving accessibility for diverse user groups. Furthermore, the survey addresses the need for robust natural language processing (NLP) systems, recommending strategies to enhance resilience against adversarial attacks. It also explores the integration of speech recognition and translation into educational feedback systems, supporting the analysis of student feedback and fostering effective communication between instructors and learners. In conclusion, this literature survey acts as a useful tool for multilingual speech transcription and translation academics, practitioners, and developers. It draws attention to the gaps in the literature that now exist and provides suggestions for future lines of inquiry, fostering more investigation and creativity in this rapidly developing field.

REFERENCES

[1] Chung, Y. A., Weng, W. H., Tong, S., & Glass, J. (2019, May). Towards unsupervised speech-to-text translation. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7170-7174). IEEE.

[2] Elakkiya, A., Surya, K. J., Venkatesh, K., & Aakash, S. (2022, December). Implementation of Speech to Text Conversion Using Hidden Markov Model. In 2022 6th International Conference on Electronics, Communication and Aerospace Technology (pp. 359-363). IEEE.

[3] Sung, T. W., Liu, J. Y., Lee, H. Y., & Lee, L. S. (2019, May). Towards end-to-end speech-to-text translation with two-pass decoding. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7175-7179). IEEE.

[4] Shi, J., Tang, Y., Lee, A., Inaguma, H., Wang, C., Pino, J., & Watanabe, S. (2023, June). Enhancing Speech-To-Speech Translation with Multiple TTS Targets. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[5] Nikolaeva, D. (2023, September). An Elementary Emulator Based on Speech-To-Text and Text-to- Speech Technologies for Educational Purposes. In 2023 XXXII International Scientific Conference Electronics (ET) (pp. 1-6). IEEE.

[6] Indurthi, S., Han, H., Lakumarapu, N. K., Lee, B., Chung, I., Kim, S., & Kim, C. (2020, May). End-end speech-to-text translation with modality agnostic meta-learning. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7904-7908). IEEE.

[7] Omar, M., Choi, S., Nyang, D., &Mohaisen, D. (2022). Robust natural language processing: Recent advances, challenges, and future directions. IEEE Access.

[8] Sunar, A. S., & Khalid, M. S. (2023). Natural Language Processing of Student's Feedback to Instructors: A Systematic Review. IEEE Transactions on Learning Technologies.

[9] Meyer, S., Tilli, P., Denisov, P., Lux, F., Koch, J., & Vu, N. T. (2023, January). Anonymizing Speech with Generative Adversarial Networks to Preserve Speaker Privacy. In 2022 IEEE Spoken Language Technology Workshop (SLT) (pp. 912-919). IEEE.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

[10] Lakdawala, B., Khan, F., Khan, A., Tomar, Y., Gupta, R., & Shaikh, A. (2018, April). Voice to Text transcription using CMU Sphinx A mobile application for healthcare organization. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 749-753). IEEE.

[11] Reddy, V. M., Vaishnavi, T., & Kumar, K. P. (2023, July). Speech-to-Text and Text-to-Speech Recognition Using Deep Learning. In 2023 2nd International Conference on Edge Computing and Applications (ICECAA) (pp. 657-666). IEEE.

[12] V. Advaith, A. Shivkumar and B. S. Sowmya Lakshmi, "Parts of Speech Tagging for Kannada and Hindi Languages using ML and DL models," 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2022, pp. 1-5, doi:10.1109/CONECCT55679.2022.9865745. keywords: {Deep learning;Machine learning algorithms;Computationalmodeling;Tagging;Linguistics;Natural language processing;Communicationstechnology;Natural Language Processing;MachineLearning;DeepLearning;Part of Speech

[13] R. V. Malage, H. Ashish, S. Hukkeri, E. Kavya and R. Jayashree, "Low Resource Speech-to-Speech Translation of English videos to Kannada with Lip-Synchronization," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 1680-1687, doi: 10.1109/ICICCS56967.2023.10142578.

[14] S. Geeta and B. L. Muralidhara, "Syllable as the basic unit for Kannada speech synthesis," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2017, pp. 1205-1208, doi: 10.1109/WiSPNET.2017.8299954.

[15] S. Hebbar, A. R. B, M. S. N, M. Supriya, N. V. G and S. L, "Named Entity Recognition Using BERT Model for Kannada Language," 2023 International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS), Manipal, India, 2023, pp. 212-216, doi: 10.1109/ICRAIS59684.2023.10367119.

[16] S. A. A T, S. Shankaran, H. M. Thrupthi and M. H R, "Natural Language Processing based Cross Lingual Summarization," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1825-1829, doi: 10.1109/ICOEI53556.2022.9776655.

[17] Danyang Liu, Ji Xu, Pengyuan Zhang, YonghongYan,A unified system for multilingual speech recognition and language identification,SpeechCommunication,Volume 127,2021,Pages 17-28,ISSN 0167-6393,https://doi.org/10.1016/j.specom.2020.12.008.

[18] Islam, Sharmi, Mustahid Hasan, and Md Ismail Jabiullah. "An Implementation of Advanced NLP for High-Quality Text-To-Speech Synthesis." Advancement of Computer Technology and its Applications 4.2, 3 (2022): 19-30.

[19] Rustam Shadiev, Yueh-Min HuangFacilitating cross-cultural understanding with learning activities supported by speech-to-text recognition and computer-aided translation,Computers&Education,Volume98,2016,Pages 130-141,ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2016.03.013.

[20] N Shikha, R Pranav, Nidhi R Singh, V Umadevi and Muzammil HussainConference: 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN), Year: 2023, Page 379, DOI: 10.1109/SPIN57001.2023.10117096