Prerequisites for the development of the system of automatic comparison of video and audio tracks by the speaker's articulation

Marsel Shakirzyanov[1], *Ruslan* Gibadullin[1,*], and *Marat* Nuriyev[2]

[1]Department of Computer Systems, Kazan National Research Technical University named after A. N. Tupolev – KAI, Kazan, Russia
[2]Department of Automated Information Processing and Control Systems, Kazan National Research Technical University named after A. N. Tupolev – KAI, Kazan, Russia

Abstract. Deep learning and reinforcement learning technologies are opening up new possibilities for the automatic matching of video and audio data. This article explores the key steps in developing such a system, from matching phonemes and lip movements to selecting appropriate machine-learning models. It also discusses the importance of getting the reward function right, the balance between exploitation and exploitation, and the complexities of collecting training data. The article emphasizes the importance of using pre-trained models and transfer learning, and the importance of correctly evaluating and interpreting results to improve the system and achieve high-quality content. The article focuses on the need to develop effective mapping quality metrics and visualization methods to fully analyze system performance and identify possible areas for improvement.

1 Introduction

The problem of video and audio track synchronization has been an issue for many years, and today we already have several technologies that can significantly simplify this process. In particular, an important role is played by the development of deep learning algorithms, which make it possible to analyze the speaker's articulation and find a correspondence between the movements of the lips and the spoken words.

Despite significant advances, there are many problems and challenges faced by the developers of such systems. In this article, we will dive into the world of modern technology and explore what factors have led to the emergence of systems for the automatic matching of video and audio tracks by speaker articulation, and what scientific breakthroughs can be catalysts for further development in this field.

[*] Corresponding author: landwatersun@mail.ru

2 Basic articulation and analysis of lip movement

2.1 Articulation and lip movement mechanisms

Articulation is the process of forming speech sounds through the movements of the articulating organs, including the lips, tongue, jaw, and larynx (Fig. 1). Different movements of the lips and other articulatory organs lead to the formation of a variety of sounds and phonemes [1].

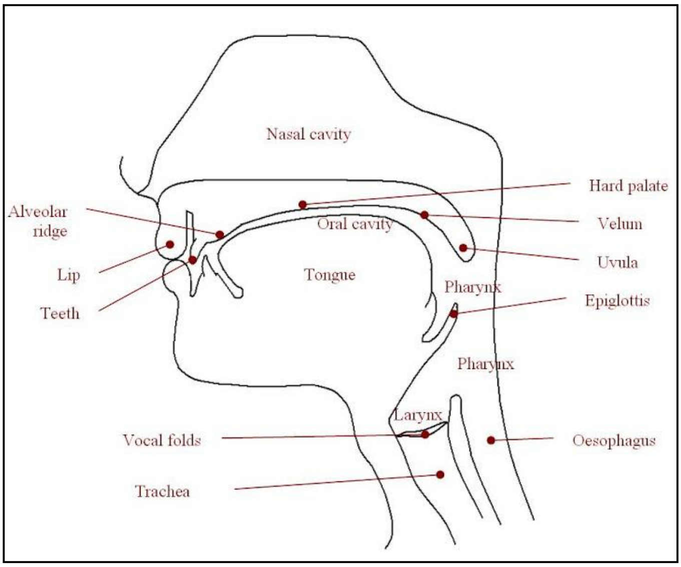

Fig. 1. Diagram of the articulating organs [2].

2.2 Methods for analyzing lip movement

There are various methods for analyzing lip movements in the context of automatic matching of video and audio tracks, including optical flow, singular point tracking, and deep learning.

1) Optical flow: Optical flow is a method that detects the motion of objects in a video by analyzing pixel brightness changes between consecutive frames (Fig. 2) [3].

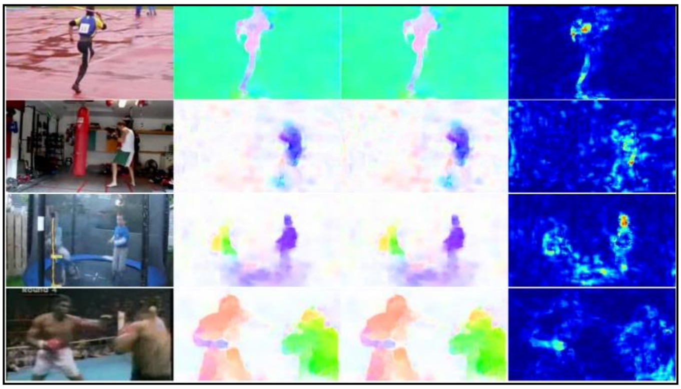

Fig. 2. Example of optical flow [4].

2) Tracking special points: Tracking special points is to track the coordinates of key points in the video, such as the corners of the lips, upper and lower lips, to analyze their movements over time (Fig. 3) [5].

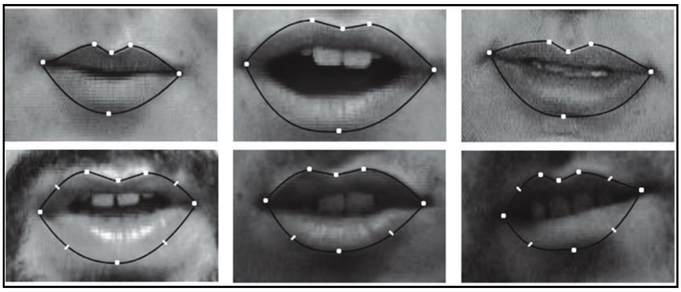

Fig. 3. Tracking special points on the example of lips [6].

3) Deep Learning: Deep learning is one of the most promising methods for analyzing lip movements, as it allows the automatic extraction of features and training of models on large amounts of data. An example of this approach is the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to analyze and compare video and audio tracks [7].

3 The application of computer vision and natural language processing

3.1 Computer vision in the analysis of lip movement

Computer vision is a field of artificial intelligence that deals with image and video processing and analysis. In the context of the automatic matching of video and audio tracks, computer vision plays an important role in analyzing lip movements and identifying articulatory features of the speaker.

1) Lip detection and tracking: One of the main tasks of computer vision in this field is the detection and tracking of lips in videos. Various algorithms such as Haar cascades, edge detection, and machine learning are used for this purpose [8].

2) Lip feature extraction: After lip detection and tracking, the next step is to extract features that describe lip movements and their articulatory characteristics. This may include measuring distances between specific lip points, analyzing lip contours, and using temporal features.

3.2 Natural Language Processing (NLP) for audio/video matching

NLP is another important area of artificial intelligence that is used to match video and audio tracks. NLP allows textual data, such as speech transcripts, to be analyzed and processed to identify patterns and match them to the speaker's articulation.

1) Phoneme analysis and audio segmentation: Phoneme analysis is the study of natural language sounds and phonemes, their classification, and segmentation. In the context of comparing video and audio tracks, phoneme analysis is used to divide the audio into distinct segments corresponding to specific sounds and articulation [9].

2) Matching phonemes and lip movements: After phonemic analysis of audio and extraction of lip features from video, the next step is to match these data. The goal is to

associate certain phonemes with corresponding lip movements, thus creating an accurate synchronization between the video and the audio tracks [10].

3) Machine learning models for audio-to-video matching: Various machine learning models, including recurrent neural networks (RNNs), hidden Markov models (HMMs), and deep neural networks (DNNs), are used to implement the mapping between phonemes and lip movements. These models are trained on large datasets containing information about video, audio tracks, and speech transcripts, and can be used to automatically match audio and video based on speaker articulation [11, 12].

3.3 Integration of computer vision and NLP in an automatic video and audio track matching system

To create an effective system for automatically matching video and audio tracks by speaker articulation, computer vision, and natural language processing must be integrated. This can be achieved by using machine learning models that process and analyze data from both domains and make decisions based on common characteristics and patterns [13, 14].

4 Artificial intelligence and deep learning in the automatic matching of video and audio tracks

4.1 Deep learning in the analysis of lip movements and audio

Deep learning is one of the most promising approaches in the field of artificial intelligence for solving the problem of automatic matching of video and audio tracks. Using deep neural networks (DNNs), large amounts of data can be processed and complex patterns in lip movements and audio signals can be identified [15, 16].

1) Convergent Neural Networks (CNNs)

Convergent Neural Networks (CNNs) are one of the main types of deep neural networks used for video analysis. CNNs are particularly effective in image processing and visual feature detection. In the context of lip motion analysis, CNNs can be used to detect and track lips and to extract features describing their articulation characteristics (Fig. 4) [17, 18].

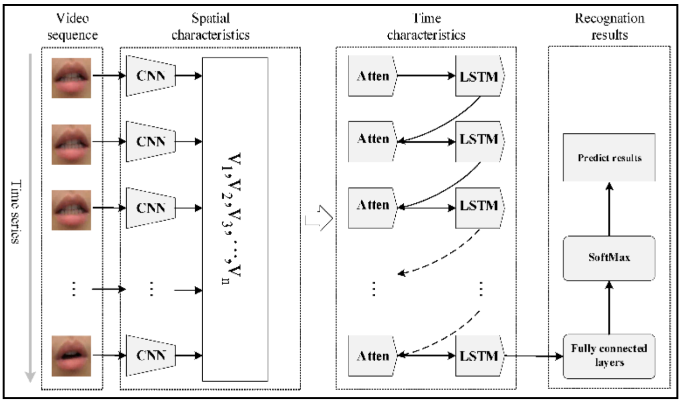

Fig. 4. Example of a convolutional neural network architecture for lip motion analysis.

2) Recurrent Neural Networks (RNN): Recurrent Neural Networks (RNNs) are another type of deep neural network that can process sequential data such as time series or audio

signals. RNNs can be used for audio analysis and segmentation, as well as for matching audio and video data based on speaker articulation (Fig. 5) [19, 20].

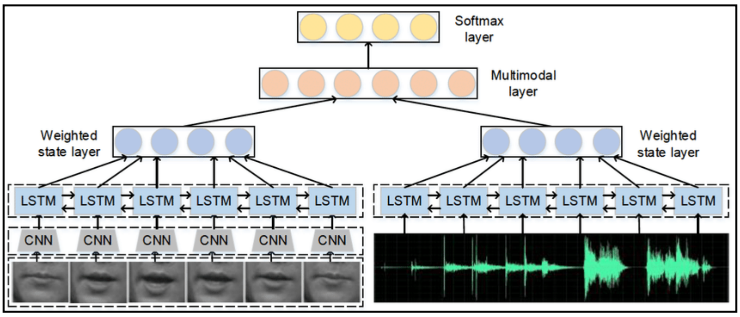

Fig. 5. Structure of the RNN multimodal model.

4.2 Reinforcement learning and synchronization optimization

Reinforcement learning is an approach in machine learning in which an agent is trained to select optimal actions based on interactions with the environment and receiving feedback in the form of rewards or penalties. In the context of automatic matching of video and audio tracks, reinforcement learning can be used to optimize the synchronization process given different constraints and application requirements [21, 22].

1) Defining the reward function: For successful applications of reinforcement learning, it is important to define a suitable reward function that evaluates the quality of matching audio and video data. The reward function should consider factors such as matching accuracy, degree of synchronization, and other criteria important to end users and applications [23, 24].

2) Exploitation and exploration: A pivotal aspect of reinforcement learning encapsulates the equilibrium between exploration, which entails probing novel potential actions and strategies, and exploitation, defined as the utilization of established and validated strategies to optimize reward. Within the context of automated alignment of video and audio tracks, this necessitates the determination of optimal parameters for machine learning models and synchronization algorithms, concurrently necessitating the adaptation of the system to diverse conditions and stipulations [25, 26].

4.3 Overcoming problems with training data

Constructing an efficacious deep learning-oriented automatic video and audio correlation system necessitates copious volumes of training data, encompassing video, audio, and speech transcripts. Nonetheless, the collection and segmentation of such data can manifest as a time-intensive and financially demanding process [27, 28].

1) Pre-trained models and transfer learning: Acceleration of the developmental process and mitigation of data collection expenses can be achieved through the application of pre-trained models and transfer learning. Pre-trained models are entities that have undergone prior training on extensive datasets and can be tailored to a specific task with minimal modifications. Transfer learning facilitates the utilization of knowledge, acquired by a model on one task, to solve other analogous tasks [29, 30].

2) Data augmentation: Data augmentation constitutes the process of generating new training instances through modifications to existing data. Within the context of automatic video and audio track alignment, augmentation could encompass the addition of noise, modulation of the pronunciation rate, rotation or scaling of images, and synthesis of novel video and audio tracks. Such strategies enhance the size and diversity of the training data,

thereby potentially elevating the performance and generalizability of deep learning models [31, 32].

4.4 Evaluation and interpretation of results

Assessing the outcomes and interpreting the performance of the automatic video and audio track alignment system constitute integral design stages. Development of evaluation metrics and methodologies is essential to gauge the system's accuracy and efficacy, taking into account the unique requisites and constraints of diverse applications [33, 34].

1) Alignment quality metrics: Alignment quality metrics may encompass measures such as the precision of phoneme and lip-movement correspondence, the degree of synchronization between audio and video data, and the proportion of correctly aligned segments. Such metrics can facilitate system developers in identifying vulnerabilities and domains necessitating further enhancement [35, 36].

2) Visualization and interpretation of results: Visualization and interpretation of system outcomes also play a crucial role in comprehending system performance and accuracy. Visualization techniques may include the exhibition of the synchronized video and audio tracks, the demonstration of algorithmic performance at various phases of the synchronization process, and the presentation of statistical data pertaining to the quality of alignment [37, 38].

5 Conclusion

In this scholarly article, we scrutinize the application of artificial intelligence (AI) and deep learning methodologies in automating the alignment of video and audio tracks. The maturation and refinement of such systems harbor immense potential to ameliorate the synchronization of audiovisual content, a critical element in a myriad of applications, including cinematography, television broadcasting, educational resources, and digital platforms [39].

The discourse encompassed various machine learning techniques instrumental in aligning video and audio tracks, such as deep neural networks, reinforcement learning, and data augmentation. The manuscript also delved into the evaluation and interpretation of results, the formulation of alignment quality metrics, and the visual depiction of data.

In summation, AI and deep learning-based automatic video and audio track alignment systems hold substantial promise to enhance the quality of audiovisual data synchronization and processing. Further evolution and research in this domain may culminate in more precise, expedited, and user-friendly systems. Such advancements can cater to a broad spectrum of users and proffer superior and more accessible content universally.

References

1. A. C. Lammert, M. I. Proctor, S. S. Narayanan, Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, 1572-1575 (2010)
2. M. Proctor, M. Coltheart, L. Ratko, T. Szalay, K. Forster, F. Cox, Memory and Cognition **49(3)**, 613-630 (2021)
3. D. J. Fleet, Y. Weiss, Handbook of mathematical models in computer vision, 237-257 (2006)
4. J. Heyman, Computers & Geosciences **128**, 11-18 (2019)

5. S. Baker, I. Matthews, International Journal of Computer Vision **60**, 221-255 (2004)
6. K. Mase, A. Pentland, Systems and Computers in Japan **22(6)**, 67-76 (1991)
7. Y. LeCun, Y. Bengio, G. Hinton, Nature **521**, 436–444 (2015)
8. P. Viola, M. Jones, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2001)
9. R. F. Gibadullin, M. Y. Perukhin, A. V. Ilin, 2021 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 398-403 (2021)
10. J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3444-3453 (2017)
11. V. A. Raikhlin, I. S. Vershinin, R. F. Gibadullin, Journal of Physics: Conference Series **2096(1)**, 012160 (2021)
12. A. Kh. Rakhmatullin, R. F. Gibadullin, Lobachevskii Journal of Mathematics **43(2)**, 473-483 (2022)
13. S. N. Cherny, R. F. Gibadullin, 2022 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 965-970 (IEEE, 2022)
14. V. A. Raikhlin, R. F. Gibadullin, I. S. Vershinin, Lobachevskii Journal of Mathematics **43(2)**, 455-462 (2022)
15. R. F. Gibadullin, M. Yu. Perukhin, B. I. Mullayanov, 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), 1-6 (IEEE, 2020)
16. R. F. Gibadullin, D. V. Lekomtsev, M. Y. Perukhin, Scientific and Technical Information Processing **48(6)**, 446-451 (2021)
17. R. F. Gibadullin, I. S. Vershinin, M. M. Volkova, 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), 1-7 (IEEE, 2020)
18. O. Soloveva, S. Solovev, R. Zaripova, F. Khamidullina, M. Tyurina, E3S Web of Conferences **258**, 11010 (2021)
19. A. Semenov, I. Yakushev, Y. Kharitonov, V. Shevchuk, E. Gracheva, S. Ilyashenko, International Journal of Technology **11(8)**, 1537-1546 (2020)
20. S. R. Khasanov, E. I. Gracheva, M. I. Toshkhodzhaeva, S. T. Dadabaev, D. S. Mirkhalikova, E3S Web of Conferences **178**, 01051 (2020)
21. V. Dovgun, S. Temerbaev, M. Chernyshov, V. Novikov, N. Boyarskaya, E. Gracheva, Energies **13(18)**, 4915 (2020)
22. A. G. Ilyin, A. S. Mahdi Khafaga, V. Yunusova, 2021 Systems of Signals Generating and Processing in the Field of on Board Communications, 1-4 (2021)
23. E. K. Vachagina, A. I. Kadyirov, K. I. Sibgatova, V. S. Yunusova, E. P. Gurmanchuk, Journal of Heat Transfer **142(11)**, 114502 (2020)
24. R. M. Shakirzyanov, A. A. Shakirzyanova, 2021 International Russian Automation Conference (RusAutoCon), 714-718 (2021)
25. M. M. Lyasheva, S. A. Lyasheva, M. P. Shleymovich, Cyber-Physical Systems: Intelligent Models and Algorithms, Cham: Springer International Publishing, 233-244 (2022)
26. M. M. Lyasheva, S. A. Lyasheva, M. P. Shleymovich, 2021 International Russian Automation Conference (RusAutoCon), 256-260 (2021)
27. M. M. Lyasheva, S. A. Lyasheva, M. P. Shleymovich, 2021 International Russian Automation Conference (RusAutoCon), 448-452 (2021)

28. V. B. Esov, A. V. Kalyashina, Russian Engineering Research **41(11)**, 1031–1034 (2021)
29. A. I. Gorunov, A. V. Kalyashina, A. A. Gabitov, Russian Engineering Research **39(7)**, 571–574 (2019)
30. A. P. Kuznetsov, H. J. Koriath, A. V. Kalyashina, T. Langer, Procedia manufacturing **21(7)**, 525-532 (2018)
31. G. Marin, D. Mendeleev, B. Osipov, and A. Akhmetshin, E3S Web of Conferences **178**, 01033 (2020)
32. Y. I. Soluyanov, A. I. Fedotov, D. Y. Soluyanov, A. R. Akhmetshin, IOP Conference Series: Materials Science and Engineering **860(1)**, 012026 (2020)
33. A. Kryukov, K. Suslov, L. Van Thao, T. D. Hung, A. Akhmetshin, Energies **15(21)**, 8249 (2022)
34. Z. Gizatullin, M. Nuriev, 2022 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 321-326 (IEEE, 2022)
35. Z. M. Gizatullin, R. M. Gizatullin, M. G. Nuriev, Journal of Communications Technology and Electronics **66(6)**, 722-726 (2021)
36. Z. M. Gizatullin, R. M. Gizatullin, M. G. Nuriev, 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 120-123 (2020)
37. K. Kulagin, M. Salikhov, R. Burnashev, 2023 International Russian Smart Industry Conference (SmartIndustryCon), 690-694 (2023)
38. R. A. Burnashev, I. A. Enikeev, A. I. Enikeev, 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), 1-4 (2020)
39. R. Burnashev, A. Enikeeva, I. F. Amer, A. Akhmedova, M. Bolsunovskaya, A. Enikeev, Lecture Notes in Networks and Systems **544** (2023)