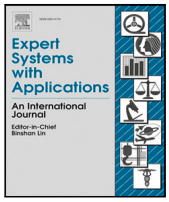
Expert Systems With Applications 252 (2024) 124159

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Audio–visual speech recognition based on regulated transformer and spatio–temporal fusion strategy for driver assistive systems

Dmitry Ryumin [a,*], Alexandr Axyonov [a], Elena Ryumina [a], Denis Ivanko [a], Alexey Kashevnik [a], Alexey Karpov [b]

[a] St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, 199178, Russia
[b] ITMO University, St. Petersburg, 191002, Russia

ARTICLE INFO

Keywords:
Audio–visual speech recognition
Spatio–temporal fusion strategy
Classifier ensemble
Transformer
Computer vision
Driver assistive systems

ABSTRACT

This article presents a research methodology for audio–visual speech recognition (AVSR) in driver assistive systems. These systems necessitate ongoing interaction with drivers while driving through voice control for safety reasons. The article introduces a novel audio–visual speech command recognition transformer (AVCRFormer) specifically designed for robust AVSR. We propose (i) a multimodal fusion strategy based on spatio–temporal fusion of audio and video feature matrices, (ii) a regulated transformer based on iterative model refinement module with multiple encoders, (iii) a classifier ensemble strategy based on multiple decoders. The spatio–temporal fusion strategy preserves contextual information of both modalities and achieves their synchronization. An iterative model refinement module can bridge the gap between acoustic and visual data by leveraging their impact on speech recognition accuracy. The proposed multi-prediction strategy demonstrates superior performance compared to traditional single-prediction strategy, showcasing the model's adaptability across diverse audio–visual contexts. The transformer proposed has achieved the highest values of speech command recognition accuracy, reaching 98.87% and 98.81% on the RUSAVIC and LRW corpora, respectively. This research has significant implications for advancing human–computer interaction. The capabilities of AVCRFormer extend beyond AVSR, making it a valuable contribution to the intersection of audio–visual processing and artificial intelligence.

1. Introduction

Audio–visual speech recognition (AVSR) represents a promising frontier in the field of automatic speech recognition (SR), offering opportunities to enhance accuracy, accessibility, adaptability, and naturalness in human–machine communication across various domains and applications.

In today's automotive landscape, the integration of advanced technologies is reshaping the driving experience. Among these technologies, SR plays a pivotal role enabling hands-free interaction with in-car systems, enhancing the driver's convenience, and promoting safety on the road. However, for traditional SR systems accuracy is challenging in diverse and dynamic environments, for example, inside a moving vehicle with prevailing background noise and visual distractions.

To address these challenges and further improve the efficacy of SR systems in automotive settings, we present an audio–visual speech command recognition transformer (AVCRFormer); which an advanced model specifically designed for robust AVSR. The proposed multi-prediction strategy demonstrates superior performance compared to traditional single-prediction strategy, showcasing the model's adaptability across diverse audio–visual (AV) contexts. The AVCRFormer architecture is based on the well-established encoder–decoder paradigm with a transformer architecture. The model's uniqueness lies in its ability to bridge the gap between acoustic and visual data, enhancing the process of recognition through an iterative refinement step.

This research provides the mathematical framework necessary to develop SR systems that are not only more efficient and resilient but contextually sensitive as well. This greater comprehension is critical for the improvements of user experiences across a variety of applications, including in-car communication and, more broadly, human–computer interaction (HCI) (Liang & Tang, 2022; Ouyang, Ma, & Tan, 2023).

The motivation behind this research lies in unlocking the full potential the AVSR technology to create intelligent and intuitive in-car communication systems that seamlessly integrate with the driver's

* Corresponding author.
E-mail addresses: ryumin.d@iias.spb.su (D. Ryumin), axyonov.a@iias.spb.su (A. Axyonov), ryumina.e@iias.spb.su (E. Ryumina), ivanko.d@iias.spb.su (D. Ivanko), alexey@iias.spb.su (A. Kashevnik), karpov@iias.spb.su (A. Karpov).

https://doi.org/10.1016/j.eswa.2024.124159
Received 27 February 2024; Received in revised form 8 April 2024; Accepted 2 May 2024
Available online 9 May 2024
0957-4174/© 2024 Elsevier Ltd. All rights reserved.

natural interaction patterns, ultimately contributing to safer, more enjoyable, and more efficient driving.

The scientific novelty of the article includes the following.

1. We have developed and presented a novel transformer-based method for AVSR based on a Transformer architecture and including a regulated cross-attention technique. The method consists of feature extraction, model initialization, and interactive model refinement steps.
2. We have developed the spatio–temporal fusion strategies of AV features and the classifier ensemble that demonstrate promising results for solving the problem of audio and video modality fusion during the command recognition.
3. We have developed and implemented the attention visualization approach that shows how different modalities (audio and video) affect overall SR accuracy. For example, in the case of noisy environments the attention shifts to the video modality, but in the case of poor lighting conditions in the vehicle or high lip occlusion it shifts to the audio one.
4. We have developed and presented a software application for driver assistive systems that is based on the AVSR method. The application integrates the proposed method, strategies, and approach that can be used for the navigation and multimedia control system in the proposed in-vehicle case study.

The remainder of this article is organized as follows: In Section 2, we provide a comprehensive review of recent advancements in audio–visual speech recognition that have recently emerged. Section 3 details the AVCRFormer architecture, covering feature extraction, model initialization, iterative refinement, and the loss function. In Section 4, we present experimental results covering encoder and decoder configurations, refinement iterations, and the gated mechanism. Additionally, we provide a comparative analysis of multi- and single-prediction strategies. Section 5 considers the implementation of the proposed method in a driver assistive system. Finally, some concluding remarks with a discussion and our vision for future research are presented in Section 6.

2. Related work

In this section, we review the state-of-the-art (SOTA) research in the field of automatic SR based on AV information processing with the focus on four different aspects: processing of audio and video signals, multimodal fusion strategies, SR in car environment, and existing AV speech corpora.

Recently, there has been a growing interest in the incorporation of combined AV information into SR systems, aiming to create more reliable and contextually aware intelligent systems (Afouras, Chung, Senior, Vinyals, & Zisserman, 2018; Chen, Hu, et al., 2023; Ma et al., 2023; Noda, Yamaguchi, Nakadai, Okuno, & Ogata, 2015; Wang, Zha, Li, Chen, & Luo, 2023).

In the field of automatic SR, neural networks (NNs) have emerged as a driving force for transformative change, spearheading a revolution in the SR technology (Bhardwaj et al., 2022; Li et al., 2022; Thomas, Kessler, & Karout, 2022; Zhang et al., 2023). Notable improvements have been made in creating SR systems with high recognition accuracy and flexibility, primarily due to the successful combination of deep learning (DL) and SR techniques. The research review the most recent state-of-the-art (SOTA) developments of SR systems featuring enhanced accuracy, efficiency, and adaptability.

2.1. Audio–visual speech recognition

Recent AVSR methods demonstrate substantial progress by transitioning from traditional handcrafted methods to such DL architectures as long-short term memory (LSTM) (Makino et al., 2019; Shashidhar, Patilkulkarni, & Puneeth, 2022), gated recurrent unit (GRU) (Liu, Li, & Yang, 2021; Petridis et al., 2018), and transformers (Burchi & Timofte, 2023; Lee, Jang, Kim, Park, & Park, 2020; Lin, Sung, Lei, Bansal, & Bertasius, 2023; Ma, Petridis, & Pantic, 2021; Song, Sun, & Li, 2022; Yang et al., 2023). This comprehensive integration enhances recognition accuracy and robustness as well as advances the AVSR performance through attention mechanisms and NN-based classifiers.

The NNs, including convolutional neural network (CNN)s, feed-forward network (FFN)s, autoencoders, and robust recurrent neural network (RNN)s, have proven to be effective in unraveling the intricacies of AV speech (Ma, Petridis, & Pantic, 2021; Makino et al., 2019; Shewalkar, Nyavanandi, & Ludwig, 2019; Zhou, Yang, Chen, Wang, & Jia, 2019). Deep NNs (DNNs) are crucial in current AVSR by assisting in feature extraction and classification. They are used in both the front-end and back-end of AVSRs, adeptly extracting features while encoding spatio and temporal information concurrently. While the significant data requirements and computational complexity of training DL models can be challenging (Michelsanti et al., 2021), transfer DL emerges as a solution. This method allows pre-trained models to refine their capabilities on new, similar tasks, thereby reducing computational and data requirements (Gerczuk, Amiriparian, Ottl, & Schuller, 2021; Zhuang et al., 2020).

CNNs excel at extracting features from audio and video data, eliminating the need for manual feature selection. This shift has revolutionized AVSR by automating the extraction of distinctive features. DL is integrated with modality fusion, enabling current AVSR systems to seamlessly combine information from audio and video modalities. Such fusion enhances recognition accuracy and robustness. Moreover, attention mechanisms and other NN-based classifiers have further improved the AVSR performance.

Transformer models have demonstrated exemplary performance in the AVSR task (Khan et al., 2022). The profound impact of Transformer models has become clearer with their scalability to very large capacity models. The other method is the Conformer Transducer presented in (Gulati et al., 2020); it blends convolutional and transformer-based NNs for audio processing, and proves to be highly effective in dealing with varying lengths of sequences. This effectiveness is achieved by employing a Conformer block consisting of two feed-forward modules flanking the multi-headed self-attention module and the convolution module, inspired by the Macaron Net structure (Lu et al., 2019).

A significant improvement in the field of SR is the joint implementation of connectionist temporal classification (CTC) and attention-based end-to-end (E2E) SR (Kim, Hori, & Watanabe, 2017). The method combines the efficiency of CTC and the attention mechanism in an E2E method. The use of multitask learning is beneficial for making SR systems stronger and faster, and for overcoming problems caused by noisy environments and long inputs. This represents an important step towards more accurate and adaptive E2E SR systems. The results encourage and offer high potential for real-world situations (Deng, Cao, Zhang, & Ma, 2021; Miao, Cheng, Gao, Zhang & and Yan, 2020; Zhang, Zhao, et al., 2020).

The authors in Dai et al. (2023) present a novel method for AVSR by decoupling the training process into two stages and introducing a visual pre-training framework based on lip shape correlation with syllables.

In work (Ma et al., 2023) the authors explore the importance of auto AVSR and the challenges associated with the accuracy of transcription labeling. The authors propose a method to leverage publicly available SR models for automatically generated transcriptions of unlabeled corpora. This method substantially expands the training set size and boosts the AVSR performance to SOTA, especially on the LRS2 and LRS3 corpora (Afouras, Chung, & Zisserman, 2018).

In Pan et al. (2022) authors research the implementation of self-supervised learning in AVSR by integrating pre-trained models trained on vast unlabeled single-modal data. The core method incorporates pre-trained models trained on extensive unlabeled data in single modalities, using the Libri-Light corpus (Kahn et al., 2020) for audio the and ImageNet corpus (Deng et al., 2009) for visual data.

Table 1
Overview of AVCR methods used in driver assistive systems. GMM refers to Gaussian Mixture Model, HMM to Hidden Markov Model, SGMM to Subspace Gaussian Mixture Model.

Work	Audio	Video	Audio–Visual	CRNN	GMM-HMM	CLSTM	RNN-GRU	CNN	DNN-HMM	SGMM-HMM
Tamazin, Gouda, and Khedr (2019)	+	−	−	−	+	−	−	−	−	−
Adeel, Gogate, and Hussain (2020)	−	−	+	−	−	+	−	−	−	−
Isobe, Tamura, Hayamizu, Gotoh, and Nose (2021)	+	+	+	−	−	−	−	+	−	−
Jaradat, Alzubaidi, and Otoom (2022)	+	−	−	−	+	−	−	−	+	+
Chuang, Wang, and Tsao (2022)	−	−	+	+	−	−	−	−	−	−
Kumar, Renuka, Rose, Wartana, et al. (2022)	+	+	+	−	−	−	+	+	−	−

In Shi, Hsu, and Mohamed (2022) self-supervised learning is also employed along with several key methods to enhance AVSR. The authors introduce the audio–visual hidden unit BERT (AV-HuBERT) framework, which uses self-supervised learning, extending the HuBERT (Hsu et al., 2021) framework for joint speech representation learning from audio and lip movement cues in video recordings. Additionally, noise augmentation is a pivotal technique employed at both pre-training and fine-tuning stages. This adds diverse noise types (ambient, babble, music, and speech noise) to the audio input, thereby improving the model's robustness in noisy environments.

These methods collectively contribute to the significant performance gains observed in AVSR, making the models highly effective and adaptable in challenging acoustic conditions. To further advance the performance of SOTA AVSRs in driving scenarios, we propose an AVSR method based on regulated transformers and a spatio–temporal fusion strategy to integrate results of audio–visual modalities. In line with the prevailing trends we have opted to employ the Transformer architecture in our method. The proposed iterative model refinement module can bridge the gap between audio and visual data by compensating for the weaknesses of unimodal information. Our transformer achieved the highest values of recognition accuracy on benchmark corpora.

2.2. Multimodal fusion strategies

Numerous studies on AVSR indicate that incorporating modalities fusion strategy can greatly improve robustness, especially in challenging acoustic environments. Several pioneering AV fusion strategies are discussed in Katsaggelos, Bahaadini, and Molina (2015), primarily categorized into feature fusion and decision fusion techniques. Feature fusion emerges as the prevailing method in AVSR (Chung, Senior, Vinyals, & Zisserman, 2017; Petridis et al., 2018). However, simple feature concatenation fails to explicitly address the reliability of both signals. Additionally, Chung et al. (2017) notes that the system tends to overly depend on the audio modality, prompting the application of a regularization technique during training, where one of the signal is randomly dropped out.

In Ma, Petridis, and Pantic (2021), introduced an end-to-end dual-encoder hybrid CTC/Attention (Kim et al., 2017) system for AVSR. This system comprises a visual encoder based on ResNet (He, Zhang, Ren, & Sun, 2015), an audio encoder based on Conformer (Gulati et al., 2020), and a multi-layer perception (MLP) module for integrating features from different modalities.

In contrast to the MLP-based fusion method, (Sterpu, Saam, & Harte, 2018) pioneered an attention-based fusion one, demonstrating that aligning features from different modalities enhances learned representations. Subsequently, numerous studies have embraced a cross-attention module along with Transformers to capture inherent alignments and complementary information within fully encoded AV representations (Sterpu et al., 2018; Wu, Li, Yang, Wu, & Qian, 2021). Some opt to directly concatenate raw speech and video sequences and utilize a shared encoder with self-attention mechanisms to learn modality alignments (Hong, Kim, Choi, & Ro, 2023; Li, Li, Wu, & Qian, 2023). In Cheng, Wang, Wang, and Li (2022), Li, Zhou, Wang, Hong, and Li (2023) researchers leveraged hidden features from various layers of audio and visual encoders to achieve more effective fusion, assuming that conducting multi-layer fusion can enhance the performance of AVSR systems.

Unlike the above AVSR methods concatenating of the feature vectors and their subsequent aggregation using NN models (Ma et al., 2023; Pan et al., 2022; Ryumin, Ivanko, & Ryumina, 2023; Shi et al., 2022) or cross-modal attention (Axyonov, Ryumin, Ivanko, Kashevnik, & Karpov, 2024; Dai et al., 2023), in our spatio–temporal modality fusion strategy implies making feature matrices uniform in size and combining them along the temporal axis. Such fusion allows preserving the contextual information of both modalities and achieving their synchronization. This integration enables the model to fuse acoustic and visual information, providing a holistic representation capturing spatio–temporal information of audio and video data. This strategy enhances the model's ability to comprehend multimodal input, facilitating feature extraction for the proposed AVCRFormer architecture.

2.3. Automatic speech recognition for driver assistive systems

Driver assistive systems with SR capabilities offer hands-free control of key vehicle functions, enhancing both the comfort and safety of the driving experience. These systems may include voice-activated navigation, enabling input of destination and adjustment of route settings without taking hands off the wheel and eyes on the road. Voice-controlled infotainment allows for control over audio playback and climate settings, often by simply requesting the desired music or temperature setting (Alvarez, Martin, Dunbar, Taiber, Wilson, & Gilbert, 2010; Biswas, Sahu, & Chandra, 2015; Zhang, Pinto, Plahl, Schuller, & Willett, 2014). Other SR systems are virtual assistants that can perform a wide range of tasks, such as sending messages, setting reminders, and managing other "to-do" functions. They are also appearing in systems that allow adjusting vehicle settings, for example changing the seat position or color of interior lighting (Braun, Mainz, Chadowitz, Pfleging, & Alt, 2019; Loh, Boey, & Hong, 2017; Mehler et al., 2015). Overall, SR aims to provide a natural and intuitive way to interact with a vehicle, reducing distractions and contributing to comfort, convenience and safety while driving.

In the specific context of a car environment, the goal is to improve SR accuracy and reliability. Engine noise, road conditions, and passenger conversations make the automotive environment noisy and dynamic, which challenges effective interaction between the assistive system and the driver (in-car communication). As a solution, researchers develop cue-based audio and visual systems that comprehend spoken instructions with precision and adjust to changing conditions. The ultimate aim of integrating AVSR into modern cars is to enhance the driving experience and promote road safety by reducing distractions and enabling drivers to interact with their vehicles more naturally and effortlessly. This could potentially lead to AVSR becoming a standard feature in cars of the next generation and represent a significant leap forward in automotive technology.

The methods used in the driver's assistive systems is summarized in Table 1. Compared to traditional AVSRs, there has been little research done on the SR systems for a vehicle, without applying transformer-based methods in them.

In Tamazin et al. (2019), the researchers introduced a modified method for calculating Power normalized cepstral coefficients (PNCC),

Table 2
Comparison of multimodal corpora. (En)glish, (Ru)ssian, (Ch)inese (include Mandarin), (Can)tonese, (Ko)rean, (Tu)rkish, (Fa)rsi. Online includes video clips of TV shows, video content on YouTube, and so on. Sent. refers to Sentences, Cmd. to Commands.

Name	# Spk.	Voc.	# Utt.	Dur., h	Resolution	FPS	View	Audio, kHz	Env.	Lang.	Content	Measure	Result, %
LRW (Chung & Zisserman, 2016)	1K+	500	539K	173	256 × 256 128 × 128	25	Front	16	Online	En	Word	Acc.	65.40
LSVSR (Shillingford et al., 2019)	1K+	127K	2.935M	3.9K	128 × 128	25	Front	–	Online	En	Sent.	WER	40.90
LRW-1000 (Yang et al., 2019)	2K+	1K	718K	508	Different	25	Multi	16	Online	Ch	Word	Acc.	38.19
LRWR (Egorov, Kostyumov, Konyk, & Kolesnikov, 2021)	135	235	117.5K	350	112 × 112	25	Multi	–	Online	Ru	Word	Acc.	61.10
RUSAVIC (Ivanko, Axyonov, et al., 2022)	20	62	12.4K	7	1920 × 1080	60	Front	48	Car	Ru	Cmd.	Acc.	87.26
CI-AVSR (Dai et al., 2022)	30	200	4.99K	8.3	640 × 480	25	Front	16	Indoor	Can	Cmd.	CER	3.48
VLRDT (Berkol, Tümer-Sivri, Pervan-Akman, Çolak, & Erdem, 2023)	–	10	2.33K	1	Different	30	Multi	–	Online	Tu	Word	Acc.	61.70
OLKAVS (Park et al., 2023)	1.1K	–	2.5M	5.75K	1920 × 1080	30	Multi	48	Lab	Ko	Sent.	CER	5.64
Arman-AV (Peymanfard et al., 2023)	1.76K	–	89K	220	224 × 224	25	Multi	–	Online	Fa	Sent.	CER	13.08
MuAViC (Anwar et al., 2023)	8K+	–	–	1.2K	Different	25	Multi	16	Online	Multi	Sent.	WER	2.30
CN-CVS (Chen, Wang, & Zheng, 2023)	2.7K	–	206.2K	308	–	25	Multi	16	Online	Ch	Sent.	–	–
MAVD (Wang, Huo, et al., 2023)	64	–	12.5K	24	1920 × 1080 RGB 640 × 576 Depth	30	Front	16	Lab	Ch	Sent.	CER	8.78

to improve SR in noisy environments, particularly in those associated with cars. The authors compared PNCC with traditional methods. Their experiments showed that in free-noise conditions, all methods performed comparably. However, when exposed to car noise at signal-to-noise ratio (SNR) 0 dB, the modified PNCC outperformed such traditional methods as Mel-frequency cepstral coefficients (MFCC) and standard PNCC.

In Adeel et al. (2020), the authors have introduced a speech enhancement framework, particularly relevant to the context of in-car communication. They leveraged DNNs, including LSTM and contextual-LSTM, to process and adapt both audio and visual features in varying automotive noise scenarios. The novelty is the development of the Enhanced Visually-derived Wiener Filter (EVWF) that significantly improves the speech quality in car noise. The work was based on data combining clean grid videos with automotive noises, which makes this system highly relevant to in-car communication applications.

The authors introduced a multi-angle VSR system with three components: view classification, feature extraction, and recognition (Isobe et al., 2021). They used a 2-Dimensional (2D)CNN for view classification and created 3DCNN models for feature extraction, taking into account the importance of various angles, especially in dynamic environments such as cars, where drivers constantly turn their heads. Recognition involved integrating features from angle-specific models and using fully-connected layers.

A SR system (Jaradat et al., 2022) has been developed to enhance driver assistive applications by minimizing driver distraction. This system uses MFCC for extracting acoustic features and the SRILM toolkit (http://www.speech.sri.com/projects/srilm/ accessed on 1 Dec. 2023) for the Bi-gram language model analysis. The results indicate that the DNN achieves a 5.17% WER for noise-free data and a 6.68% WER for noisy data.

In Chuang et al. (2022) the authors describe a novel system aimed at enhancing AV speech under challenging conditions. They use a three-stage approach, involving data pre-processing, a convolutional recurrent neural network (CRNN)-based audio–visual speech enhancement (AVSE) stage, and data reconstruction, and address real-world problems such as the computational cost of processing visual data, AV asynchrony, and low-quality visual data. In particular, under challenging conditions, the model achieves a remarkable accuracy across different scenarios. The results underscore the effectiveness of their AVSR method, making it a valuable contribution to the field.

In the context of assisting individuals with hearing impairments, an AVSR system focused on accommodating the challenging in-car environment was developed (Kumar et al., 2022). This system leverages both audio and visual modalities to ensure effective communication.

In general, driver SR systems are gaining more popularity. They are designed to enhance the accuracy and robustness of SR for common driver requests in real-world driving scenarios. Given the inherent challenges of SR in driving conditions, including ambient noise, head movements, varying poses, distance from recording devices, and lighting conditions, the researchers leverage AV information to improve recognition results. Despite of the variety of methods used, it is worth noting that transformers have not previously been used in such systems. In this work, we improve existing driver SR systems by using the SOTA AVCRFormer methodology.

Due to DL techniques, including convolutional and long short-term memory networks, AVSR has significantly advanced. These advancements have led to improved recognition accuracy and robustness, with practical implications in noise-robust SR, accessibility for individuals with hearing impairments. Recent studies have achieved the SOTA performance by addressing challenges related to modality variations and cross-modal alignment. In the automotive context, the AVSR systems

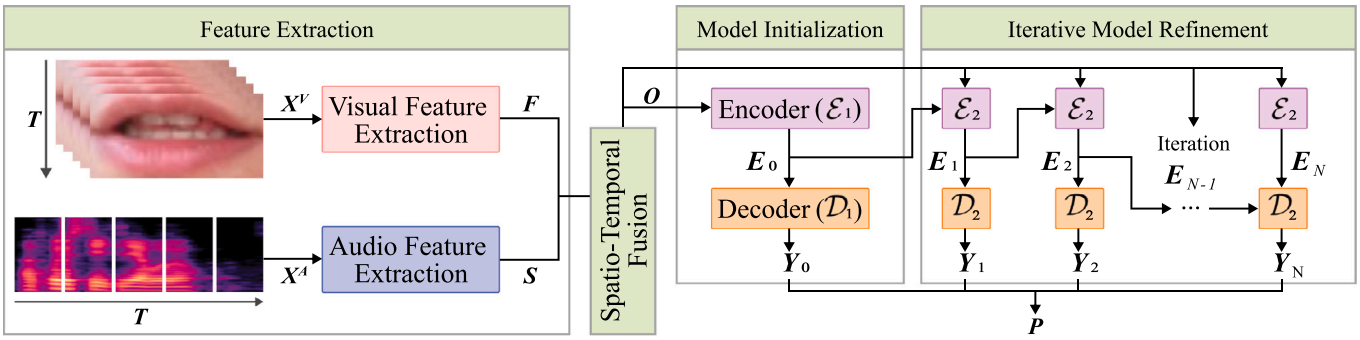


Fig. 1. Pipeline of the AVCRFormer architecture.

aim to enhance SR accuracy and improve the overall driving experience. The proposed techniques like EVWF and multi-angle visual SR have demonstrated significant improvements in challenging conditions, making them valuable for reducing distractions and enhancing road safety. Therefore, AVSR promises various practical applications and continues to evolve with ongoing research and development efforts.

2.4. Audio–visual speech corpora

AV data processing has become one of the fastest-evolving research areas within artificial intelligence (AI) and CV. The availability of high quality corpora for the study of SR and AV signals has become a major factor in this development. This section reviews the most recent corpora designed for AVSR, presented in various languages, including multi-lingual corpora. The research also includes the baselines achieved by researchers. Table 2 shows the distinctive characteristics of recent corpora.

Based on the conducted analysis, we chose two AV corpora for model training and experimental evaluation. In the context of this research, the selection of the Russian Audio–Visual Speech in Cars (RUSAVIC) corpus (Ivanko, Axyonov, et al., 2022) is of great importance. Distinguished by its recordings made in authentic driving scenarios, RUSAVIC proves to be a versatile and carefully controlled corpus, offering a spectrum of SNRs influenced by dynamic elements such as window opening/closing and car motion. This corpus matches the focus of the research, encompassing driver commands tailored to control car navigation and multimedia systems. The inclusion of RUSAVIC in the experiments enhances the relevance and authenticity of the research and makes a significant contribution to the advancement of the AVSR research in the specific context of the automotive environment.

At the same time, we also consider the LRW corpus (Chung & Zisserman, 2016) for our experiments. Renowned for its extensive usage in the AVSR domain, LRW serves as a benchmark corpus in our research. While various corpora are used across a broader AVSR context, we opt for RUSAVIC and LRW due to their unique characteristics that align well with the specific goals and evaluation criteria of our research.

The comparison of corpora reveals significant differences, not only in recording characteristics such as the number of speakers, duration, and resolution, but also in the languages and tasks they are intended for. It is worth noting that corpora are primarily collected for continuous SR tasks, labeled as *"Sent."* in Table 2. However, these corpora may not be the optimal choice for driver assistive systems. The table includes other corpora, labeled *"Word"*, which are collected for single word recognition tasks. These corpora may be more appropriate for use in driver assistive systems, but they may contain excessive vocabulary. While the corpora are labeled as *"Cmd."*, they are specifically developed for driver assistive systems as they are collected based on frequently encountered driver requests. As mentioned earlier, for our research, we have chosen the RUSAVIC and LRW (Chung & Zisserman, 2016) corpora.

3. Methodology

Driver assistive systems demand ongoing interaction with the human driver. However, while on the road, the driver is unable to use their hands to interact with the system. Besides, vehicle environment has different kind of noise: sound noise and visual noise due to lighting conditions. We proposed an AVSR method based on regulated transformers as well as spatio–temporal fusion strategy to integrate audio and video modalities results. The method allows distributing the model's attention depending on the informativeness of the modalities, thus partially solving the problem of noisy audio or video signals. In this section, we present the overall architecture of our AVCRFormer and then discuss each component separately. Finally, we outline the training aim and inference strategy for AVCRFormer in the context of the intelligent driver assistant.

3.1. AVCRFormer method overview

In line with the prevailing trends in AVSR methodology, we have opted to employ the Transformer architecture in our proposed method. Fig. 1 offers a pipeline of the AVCRFormer architecture, encompassing four primary modules:

1. Feature extraction;
2. Spatio–temporal fusion;
3. Model initialization;
4. Iterative model refinement.

The foundation of the AVCRFormer method lies in obtaining a representation of the mapping function $f : \boldsymbol{X^A}, \boldsymbol{X^V} \rightarrow \boldsymbol{P}$ for each AV recording of drivers in cars. Where $\boldsymbol{X^A}$ and $\boldsymbol{X^V}$ refer to audio and video input signals, $\boldsymbol{P}$ to class prediction. The audio $\boldsymbol{X^A} = \{\boldsymbol{x}_t^A\}_{t=1}^T$ and video $\boldsymbol{X^V} = \{\boldsymbol{x}_t^V\}_{t=1}^T$ signals are considered as input within AVCRFormer where T refers to the audio spectrograms or the video frames. Then, two different feature extractors are used. The first extractor transforms visual input $\boldsymbol{X^V}$ into feature matrix $\boldsymbol{F} = \{\boldsymbol{f}_t^V\}_{t=1}^T$, while the second extractor transforms acoustic input $\boldsymbol{X^A}$ into feature matrix $\boldsymbol{S} = \{\boldsymbol{f}_t^A\}_{t=1}^T$. These sequential feature matrices together form the spatio–temporal features (STF). After the feature extraction module, the STF pass through a multimodal fusion module based on the concatenation of features along the temporal axis. The output of the module is matrix $\boldsymbol{O}$. The model initialization module then uses the $\mathcal{E}_1$ encoder and D_1 decoder to generate initial data representation $\boldsymbol{E}_0$ and initial recognition $\boldsymbol{Y}_0$ from matrix $\boldsymbol{O}$.

To bridge the divide between acoustic and visual signals and contextual representation, we enhance the initial process of data representation and recognition by incorporating an iterative model refinement module with the multiple feature representations and the classifier ensemble. Unlike the previous module, where the encoder only analyzes matrix $\boldsymbol{O}$ and calculates a new matrix based on the self-attention, the encoder of this module analyzes the cross-attention of two matrices: the matrix $\boldsymbol{O}$ and the matrix obtained at the previous step.

The iterative model refinement module, involving encoder $\mathcal{E}_2$ and a decoder D_2, takes initial data representation $\boldsymbol{E}_0$ as input and generates data representation $\boldsymbol{E}_N$ and probability prediction $\boldsymbol{Y}_N$ for the current step by integrating it with initial AV representation $\boldsymbol{O}$. After the N iterations of refinement, we obtain a sequence of data representations $\boldsymbol{E} = \{\boldsymbol{E}_0, \boldsymbol{E}_1, \ldots, \boldsymbol{E}_N\}$ and their corresponding probability prediction sequence $\boldsymbol{Y} = \{\boldsymbol{Y}_0, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$, therefore, $\boldsymbol{E} \rightarrow \boldsymbol{Y}$. The final class prediction ($\boldsymbol{P}$) is computed using the argmax operation with the elements of the average of all probability prediction vectors in $\boldsymbol{Y}$:

$$\boldsymbol{P} = argmax(\frac{1}{N}\sum_{n=0}^{N-1}\boldsymbol{Y}_n), \tag{1}$$

where, N represents the total number of refinement iterations. The averaging mechanism helps enhance the stability and robustness of the class prediction based on the refined information from all iterations.

Hence, the feature extraction module extracts and combines features from audio and video inputs. The model initialization module is responsible for the initial representation (aggregation) of AV features and the initial recognition. The iterative model refinement module (at each N iteration) takes the combined matrix of AV features and the feature representation obtained in the previous iteration as an input and produces a refined feature representation and recognition as an output. Due to its iteration, the feature representation is constantly updated and new (most informative) features are searched for in the original matrix of AV features. Based on the informativeness of the features, the recognition changes at each iteration, so that the final class prediction is based on the classifier ensemble. Unlike the AVSR method in (Chen, Hu, et al., 2023), which employs multiple models to create a classifier ensemble, our method uses a single holistic model.

The interactive model refinement module with multiple feature representations and the classifier ensemble allows implementing a regulated transformer. The attention visualization of features presented in Section 4.8 demonstrates the effectiveness of the proposed regulated transformer.

3.2. Feature extraction

3.2.1. Audio feature extraction

The audio feature extractor converts a series of audio spectrograms $\boldsymbol{X}^{\boldsymbol{A}} \in \mathbb{R}^{S^A} \in \mathbb{R}^{T\times B\times D}$, where S^A refers to the audio samples, T to the number of spectrograms, B to Mel-frequency and D to the time duration, into spectrogram features $\boldsymbol{S} \in \mathbb{R}^{T\times C}$, with C representing the feature size. The pre-trained model proposed in Ryumin et al. (2023) is used as a feature extractor. The model based on ResNet18 extracts 512 deep features for each spectrogram calculated by the method proposed in Ryumin et al. (2023). The spectrograms from the audio signal are facilitated using Librosa (https://librosa.org/ accessed on 1 Dec. 2023) (McFee et al., 2015). The processing of audio signal through this extractor results in STF ($\boldsymbol{S}$), as expressed by the following equation:

$$\boldsymbol{S} = AudioFE\left(SpecE\left(\{\boldsymbol{x}_t^A\}_{t=1}^T\right)\right). \tag{2}$$

The `AudioFE` feature extractor accepts segmented spectrograms as inputs and extracts features for each segment. The sequence of features obtained from a series of spectrograms forms STF ($\boldsymbol{S}$). In our research, the feature extractor is applied to enhance the analysis of audio features, contributing to a more comprehensive understanding of the AV context in the proposed AVCRFormer architecture.

3.2.2. Visual feature extraction

The visual feature extractor transforms sequence of video frames $\boldsymbol{X}^{\boldsymbol{V}} \in \mathbb{R}^{T\times H\times W\times 3}$, where T refers to the number of frames, W and H to width and height, into visual features $\boldsymbol{F} \in \mathbb{R}^{T\times C}$, with C representing the feature size. The pre-trained model proposed in Ryumin et al. (2023) is used as a feature extractor. We use our model because there are no pre-trained models with open-source code that have performed better on the LRW corpus. In addition, our pre-trained model that combines ResNet18 and BiLSTM achieves the highest performance for automatic lip-reading tasks. It is supported by recent studies (https://paperswithcode.com/sota/lipreading-on-lip-reading-in-the-wild accessed on 28 Mar. 2024) in which ResNet18 model is consistently ranked among the top models for lip-reading tasks due to its competitive recognition accuracy and computational efficiency. Since the model was trained on the LRW corpus for AVSR, it allows extracting features for 29 frames. If a video contains fewer than 29 frames, all frames are duplicated to meet the required number. If there are more than 29 frames, the video is segmented into 29-frame segments. Then, 512 deep features are extracted for each frame. Processing video frames through this extractor yields STF ($\boldsymbol{F}$), as illustrated by the following Eq. (3):

$$\boldsymbol{F} = \text{VideoFE}\left(\{\boldsymbol{x}_t^V\}_{t=1}^T\right). \tag{3}$$

The `VideoFE` feature extractor accepts segmented frames as inputs and extracts features for each segment. The sequence of features obtained from a segmented frames forms STF ($\boldsymbol{F}$). The particular feature extractor is contingent on the task demands; in our research, it involves analyzing drivers' lip regions during their visual utterance of words while driving. The lip regions are extracted with MediaPipe, a cross-platform open-source environment (https://google.github.io/mediapipe/ accessed on 1 Dec. 2023). Therefore, the feature extractor is essential in converting raw video frames into a compressed representation, making the foundation for further analysis within the context of our proposed AVCRFormer architecture.

3.3. Spatio–temporal fusion strategy

The proposed strategy allows integrating audio and video modalities at the feature level for AVCR. Within the strategy, we implement audio and video signal pre-processing and signal segmentation for feature extraction. Then, we propose to fuse audio features with visual ones by the temporal axis and use the regulated transformer model to decode multimodal speech. The use of modality fusion by the temporal axis allows solving the problem of synchronizing audio and video data, and also enables the transformer to analyze both modalities in one pass, regulating their feature informativeness.

To construct a holistic model that includes both audio and visual features, we combine the outputs of the two feature extractors. The $\boldsymbol{S}$ feature matrix, from `AudioFE`, operates on audio spectrograms $\{\boldsymbol{x}_t^A\}_{t=1}^T$ to extract STF. Simultaneously, the $\boldsymbol{F}$ feature matrix, derived from `VideoFE`, transforms video frames $\{\boldsymbol{x}_t^V\}_{t=1}^T$ into STF.

Unlike existing AVSR methods that involve a concatenation of the feature vectors and their subsequent aggregation using NN models (Ma et al., 2023; Pan et al., 2022; Ryumin et al., 2023; Shi et al., 2022) or a cross-modal attention of the features from different modalities (Axyonov et al., 2024; Dai et al., 2023), we make STF matrices uniform in size and combine them along the temporal axis. Such fusion allows preserving the contextual information of both modalities and achieving their synchronization.

As mentioned above, the visual feature matrix has a size of 29×512, and in practice, the number of frames in a video is not fixed. We segment the video into 29 frames and extract features for all segments. To obtain the number of spectrograms equal to the number of frames in the video, we segment the audio using a fixed step. Thus, feature matrices from both modalities are obtained in the same size, containing STF information. Subsequently, the features are concatenated along the temporal axis and input to the model initialization encoder, $\mathcal{E}_1$, and the iterative model refinement encoder, $\mathcal{E}_2$.

This integration enables the model to fuse acoustic and visual information, providing a holistic representation capturing STF of audio and video data. This method enhances the model's ability to comprehend multimodal input, facilitating comprehensive feature extraction

for the proposed AVCRFormer architecture. To achieve a synergistic representation, we introduce a `SpatioTemporalFusion` strategy, denoted as $\boldsymbol{O}$:

$$\boldsymbol{O} = \text{SpatioTemporalFusion}\left(\boldsymbol{S}, \boldsymbol{F}\right), \tag{4}$$

where, $\boldsymbol{S}$ and $\boldsymbol{F}$ represent audio and video STF, respectively.

The `SpatioTemporalFusion` strategy merges the $\boldsymbol{S}$ and $\boldsymbol{F}$ features according to Algorithm 1. The input signals are first segmented into segment with fixed steps ($step^A$ and $step^V$). The number of segments depends on the duration of the signal. Then, features are extracted using AudioFE and VideoFE extractors for each audio and video segment. The resulting features are recorded in the corresponding $\boldsymbol{S}$ and $\boldsymbol{F}$ matrices. If the size of the resulting matrices is less than required size $\boldsymbol{T^R}$, padding is applied based on the mean feature values of the matrices. Finally, the matrices are concatenated along the temporal axis. The resulting fused representation, $\boldsymbol{O}$, captures combined information from both modalities and improves the overall understanding of the AV context.

Algorithm 1 SpatioTemporalFusion

procedure SpatioTemporalFusion(X^A,X^V,T^R)
  $S^A \leftarrow \text{shape}(X^A)$ ▷ S^A is the audio samples
  $T, _, _, _ \leftarrow \text{shape}(X^V)$ ▷ T is the video frames
  $\boldsymbol{F} \leftarrow \text{zeros}(T, C)$ ▷ $\boldsymbol{F}$ is the video STF
  $\boldsymbol{S} \leftarrow \text{zeros}(T, C)$ ▷ $\boldsymbol{S}$ is the audio STF
  $step^A \leftarrow S^A$ **div** T ▷ C is the number of features
  $step^V \leftarrow 29$
  AudioFeatureExtractor (AudioFE):
  for t, $start$ **from** enumerate(0 **to** $S^A - 1$ **step** $step^A$) **do**
    $end \leftarrow start + step^A - 1$
    if $end \geq S^A$ **then**
      $end \leftarrow S^A - 1$
    end if
    $X_t^A \leftarrow \text{SpecE}(X^A[start : end])$ ▷ $t \in T$
    $\boldsymbol{S}[t, ...] \leftarrow \text{AudioFE}(X_t^A)$
  end for
  VideoFeatureExtractor (VideoFE):
  for $start$ **from** 0 **to** $T - 1$ **step** $step^V$ **do**
    $end \leftarrow start + step^V - 1$
    if $end \geq T$ **then**
      $end \leftarrow T - 1$
    end if
    $\boldsymbol{F}[start : end, ...] \leftarrow \text{VideoFE}(X^V[start : end])$
  end for
  PaddingMatrices:
  $\boldsymbol{S} \leftarrow \text{padding}(\boldsymbol{S}, \text{need_height} = T^R, \text{mode} = mean)$
  $\boldsymbol{F} \leftarrow \text{padding}(\boldsymbol{F}, \text{need_height} = T^R, \text{mode} = mean)$
  ConcatenateMatrices:
  $\boldsymbol{O} \leftarrow \begin{bmatrix} S_{T^R \times C} \\ F_{T^R \times C} \end{bmatrix}$
  return $\boldsymbol{O}$
end procedure

3.4. Model initialization

Following the concatenation of audio and visual features, the combined feature matrix is processed in the model initialization module of AVCRFormer, which encompasses an encoder–decoder architecture. The encoder, denoted as $\mathcal{E}_1$, uses the concatenated features to create an encoded representation, while the decoder, denoted as $\mathcal{D}_1$, reconstructs the encoded information. This initialization step is crucial to establish the model's foundation.

Matrix $\boldsymbol{O} \in \mathbb{R}^{2T \times C}$ represents the concatenated AV features, where $2T$ is the number of feature vectors, and C is the number of features. Encoder $\mathcal{E}_1$ processes $\boldsymbol{O}$ to obtain encoded representation $\boldsymbol{E}_0 = \left\{\boldsymbol{e}_t^0\right\}_{t=1}^{2T} \in \mathbb{R}^{2T \times C}$, and decoder $\mathcal{D}_1$ reconstructs this representation, yielding initial probability prediction $\boldsymbol{Y}_0$. This process forms the initial step in the AVCRFormer method.

Encoder $\mathcal{E}_1$ is composed of several layers (see Fig. 2), where $\boldsymbol{E}_0$ denotes the output. The hidden state is computed as:

$$\widetilde{\boldsymbol{E}}_0 = \alpha \cdot \boldsymbol{O} + (1 - \alpha) \cdot \text{attn}_s\left(\boldsymbol{O}\right), \tag{5}$$

where attn_s denotes the self-attention layer in Transformer; $\alpha \in [0, 1]$ is a gated mechanism hyperparameter that weights the importance of previous prototypes during training and inference. To further fuse and refine the prototype, $\widetilde{\boldsymbol{E}}_0$ is linked with a fully-connected layers FCL(·) using the residual connection. The final output of encoder $\mathcal{E}_1$ is formulated as:

$$\boldsymbol{E}_0 = \text{FCL}\left(\text{FCL}\left(\text{LN}\left(\widetilde{\boldsymbol{E}}_0\right)\right) + \text{LN}\left(\widetilde{\boldsymbol{E}}_0\right)\right), \tag{6}$$

where LN(·) is the layer normalization operation. Probability prediction vector $\boldsymbol{Y}_0$ is obtained by feeding the output of encoder $\mathcal{E}_1$ into the input of decoder $\mathcal{D}_1$ (see Fig. 2). The probability prediction vector in the recognizer is computed as shown in Eq. (7):

$$\boldsymbol{h}_0 = \text{FCL}\left(\boldsymbol{E}_0\right), \tag{7}$$

$$\boldsymbol{Y}_0 = \text{FCL}\left(\boldsymbol{h}_0\right) = \text{softmax}\left(\boldsymbol{h}_0 \cdot \boldsymbol{W}_0\right), \tag{8}$$

where, $\boldsymbol{h}_0 \in \mathbb{R}^H$ represents the output of the final layer, H is a hyperparameter customized during training and used in inference, and $\boldsymbol{W} \in \mathbb{R}^{H \times M}$ denotes a linear mapping to project hidden state $\boldsymbol{h}_0$ into the probability prediction for M classes. The probability is obtained by applying the $\text{softmax}(\cdot)$ function.

Thus, through the self-attention layer, residual connections, and fully-connected layers, the model initialization module forms an initial representation of features with adapted importance of input sequence elements, leading to initial recognition.

3.5. Iterative model refinement

Once the initial prototype, denoted as $\boldsymbol{E}_0$, is obtained through the model initialization module, it is combined with the initial AV representation, $\boldsymbol{O}$, and fed into the iterative model refinement module. The iterative model refinement module consists of two intertwined subprocesses: iterative prototype aggregation within the $\mathcal{E}_2$ encoder and recognition generation within the $\mathcal{D}_2$ decoder.

Encoder $\mathcal{E}_2$ is composed of several layers (see Fig. 2), where $\boldsymbol{E}_n$ denotes the output of the nth iteration. The hidden state is computed as:

$$\widetilde{\boldsymbol{E}}_n = \alpha \cdot \text{attn}_s\left(\boldsymbol{O}\right) + (1 - \alpha) \cdot \text{attn}_c\left(\boldsymbol{E}_n, \boldsymbol{O}\right), \tag{9}$$

where attn_c denotes the cross-attention layer used in Transformer. To further fuse and refine the prototype, $\widetilde{\boldsymbol{E}}_n$ is linked with fully-connected layers FCL(·) using the residual connection. The final output of encoder $\mathcal{E}_2$ at the nth iteration is formulated as:

$$\boldsymbol{E}_n = \text{FCL}\left(\text{FCL}\left(\text{LN}\left(\widetilde{\boldsymbol{E}}_n + \boldsymbol{O}\right)\right) + \text{LN}\left(\widetilde{\boldsymbol{E}}_n + \boldsymbol{O}\right)\right), \tag{10}$$

Simultaneously, refined prototype $\boldsymbol{E}_n$ is used in the decoder $\mathcal{D}_2$ (see Fig. 2) to generate probability prediction $\boldsymbol{Y}_n$. Probability prediction $\boldsymbol{Y}_n$ at the nth iteration is calculated as $\boldsymbol{Y}_0$:

$$\boldsymbol{h}_n = \text{FCL}\left(\boldsymbol{E}_n\right), \tag{11}$$

$$\boldsymbol{Y}_n = \text{FCL}\left(\boldsymbol{h}_n\right) = \text{softmax}\left(\boldsymbol{h}_n \cdot \boldsymbol{W}_n\right), \tag{12}$$

here, $\boldsymbol{h}_n \in \mathbb{R}^H$ represents the output of the final layer.

The iterative model refinement module, unlike the model initialization module, includes an additional layer of cross-modal attention. The module encoders take the original feature representation and the previous feature representation as input. This allows re-adapting the most informative features from the original feature representation and taking into account the importance of the features highlighted in the previous step. This process is repeated the N iterations, each time obtaining an update feature representation and a new recognition.

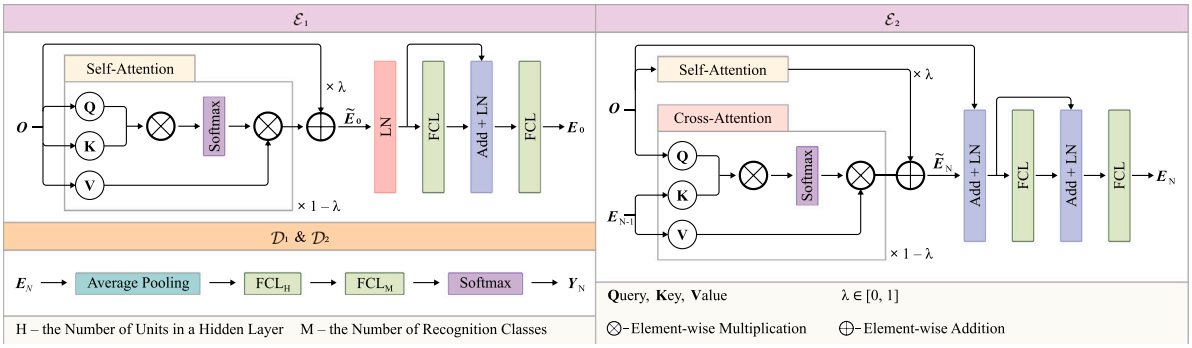


Fig. 2. Architectures of the regulated transformer components in AVCRFormer. $\mathcal{E}_1$ and $\mathcal{E}_2$ refer to encoders, $\mathcal{D}_1$ and $\mathcal{D}_2$ to decoders.

3.6. Loss function

Our method generates the recognition sequence based on the conditional probability provided by decoders D_1 and D_2. The cross-entropy loss (CE-loss) function is used to measure the difference between the ground truth labels and the decoder outputs for all training samples. The primary aim while training is to maximize the log-likelihood, equivalent to minimizing the CE-loss, defined as:

$$L_{CE,n} = -\log Y_n(\hat{\boldsymbol{Y}} \mid \boldsymbol{O}), \tag{13}$$

where $\hat{\boldsymbol{Y}}$ represents the ground truth annotation, Y_n is each probability prediction. This loss is applied to the output of the model initialization module and the Nth output of the iterative model refinement module. The overall loss function for our AV command classification system is:

$$L = L_{CE,0} + L_{CE,N}, \tag{14}$$

where $L_{CE,0}$ is CE-losses at the output of the model initialization module and $L_{CE,N}$ the Nth output of the iterative model refinement module. Initially, guided by the CE-loss, we train the initial generation module until convergence. Subsequently, the complete loss function is employed for E2E optimization of the system.

4. Experimental results

4.1. Experimental corpora

To evaluate the effectiveness of the proposed AVCRFormer, we conducted extensive experiments on two different corpora: RUSAVIC (Ivanko, Axyonov, et al., 2022) and LRW (Chung & Zisserman, 2016). The primary aim of these experiments was to achieve the SOTA accuracy in AV utterance recognition for both corpora, with a special focus on the RUSAVIC corpus. This focus is due to the fact that the RUSAVIC corpus contains essential commands relevant to the AV driver assistive system. Meanwhile, LRW served as a benchmark for a broader comparison within the context of the overall SOTA. See Table 3 for the main characteristics of the RUSAVIC and Table 4 for LRW corpora.

To comply with the experimental setup, we use data split similar to previous studies (Axyonov et al., 2024; Chung & Zisserman, 2016). The number of frames (T) varies in the RUSAVIC corpus (see Table 3) because that both single- and multi-word commands are used in this corpus. We use the maximum number of frames $T^R = 120$, as in Axyonov et al. (2024). In case the number of frames is less, we apply padding by mean feature values (see Algorithm 1). The fixed step ($step^A$) for spectrogram segmentation is set to 16 ms by 16 kHz sample rate. For the LRW corpus, T is constant and equals $T^R = 29$ and $step^A = 34$ ms by 16 kHz sample rate.

Table 3
RUSAVIC corpus characteristics.

Set	# Classes	# Samples for Each Class	# Frames
Train		140	
Val	62 (utterances)	30	[21, 172]
Test		30	

Table 4
LRW corpus characteristics.

Set	# Classes	# Samples for Each Class	# Frames
Train		800–1000	
Val	500 (words)	50	29
Test		50	

4.2. Experimental setup

During training, we meticulously fine-tuned several critical parameters, tailoring the architecture to the specific needs of our task. The primary optimized parameters:

- **Encoder and Decoder Configurations:** We varied the structure of both the encoder and decoder, tuning the number of fully-connected layers in both and adjusting the units in the hidden fully-connected layer of the decoder. This was done to find an optimal balance for feature representation and prediction generation.
- **Number of Iterations (N):** We explored the impact of varying the number of iterations (N) in the iterative model refinement to understand how many refinement steps were necessary for the model to converge to optimal performance.
- **Gated Mechanism (α):** The hyperparameter, α, controlling the weight between original prototypes and gated-based prototypes during training and inference, was systematically tuned to optimize the contribution of previous iterations to the final prediction.
- **Training Parameters:** The choice of learning rate (LR), LR decay schedule, and optimizer were considered and fine-tuned to optimize the convergence speed and overall performance. To identify the most effective optimization strategy for our model, we explored popular optimizers, including Adam (Kingma & Ba, 2014), AdamW (Loshchilov & Hutter, 2019), stochastic gradient descent (SGD) (Ruder, 2016), and the Lion optimizer (Chen, Liang, et al., 2023).
- **Multi-Prediction Strategy vs. Single-Prediction Strategy:** We investigated whether generating multi-predictions at each iteration provided any benefit compared to single-predictions. This involved assessing the trade-off between increased computation and potential improvement in model performance.

Table 5
Impact of the Encoder and Decoder configurations on recognition accuracy (%) on the RUSAVIC and LRW corpora.

Experiment	Encoder	Decoder		Accuracy (%)			
	# FCLs	# FCLs	# Units	RUSAVIC		LRW	
				Val	Test	Val	Test
Baseline 1 (Default)	1	1	128	85.38	85.22	86.63	85.72
Config 1	2	1	128	89.89	89.52	86.08	85.96
Config 2	3	1	128	88.12	88.71	86.89	86.31
Config 3	2	2	128	90.43	89.09	88.21	87.69
Config 4	2	3	128	89.14	88.39	88.30	88.65
Config 5	2	0	–	88.66	88.93	83.33	83.07
Config 6	2	1	256	90.91	90.22	88.47	87.93
Config 7 (Baseline 2)	2	1	512	**90.27**	**90.75**	88.73	88.61
Config 8	2	1	1024	89.84	90.11	**89.31**	**89.52**
Config 9	2	1	2048	88.23	87.42	89.01	88.86

This section details the experiments and provides insights into the model's training dynamics, convergence, and performance accuracy on these two corpora.

The following default parameters were set and used in all the experiments:

- A single fully-connected layer encoder.
- A single layer decoder with 128 hidden units.
- A single model for iterative refinement.
- Without α hyperparameter.
- The Adam optimizer with 1×10^{-3} LR and constant LR over 100 epochs.
- A single-prediction strategy.

4.3. Optimizing encoder and decoder configurations

We systematically optimized the AVCRFormer method by adapting the architecture of its encoder and decoder components. Our aim was to identify the optimal configuration that improves feature representation and sequence generation for improved accuracy in AV utterance recognition on both the RUSAVIC and LRW corpora. We iteratively varied the number of fully-connected layers and units, carefully balancing model complexity and recognition accuracy. Previous studies (Ryumina, Dresvyanskiy, & Karpov, 2022) showed that effective fine-tuning of models varies depending on the experimental corpus. Therefore, we do not use traditional grid search to find the optimal combination of parameters. We varied only one parameter in each experiment to obtain more unbiased accuracy on two different corpora. Table 5 provides a detailed representation of the recognition accuracy for different configurations. First, we determined the optimal number of fully-connected layers in encoders. Next, we determined the optimal number of fully-connected layers in decoders. Finally, we determined the optimal number of hidden units in the fully-connected layer of the decoders.

The experiments revealed that increasing the complexity of the encoder and decoder components generally improved accuracy, as shown in Table 5. In particular, the search for the best configuration, exemplified by "Config 7", achieved the highest accuracy at 90.75% on the RUSAVIC corpus test set. Additional complexity adjustments, such as the inclusion of 1024 units in "Config 8", resulted in only a slight decrease in accuracy.

This trend remained consistent for the LRW corpus, which is characterized by a more extensive set of classes and data (see Tables 3 and 4). While "Config 8" achieved slightly lower accuracy than in RUSAVIC, it demonstrated robust performance on both the Val (89.31%) and Test (89.52%) sets.

The results highlighted the nuanced relationship between encoder–decoder complexity and recognition accuracy across different corpora. Thus, to optimize the AVCRFormer method for accurate AV word or utterance recognition, careful configuration selection is critical, especially when dealing with varying class sizes and corpus complexity.

Table 6
Impact of the number of iterations (N) on recognition accuracy (%) on the RUSAVIC and LRW corpora.

Experiment	# Iterations (N)	Accuracy (%)			
		RUSAVIC		LRW	
		Val	Test	Val	Test
Baseline 2 (Default)	1	90.27	90.75	88.73	88.61
Config 1	2	91.88	91.13	91.13	90.64
Config 2	3	92.53	91.72	91.59	90.92
Config 3 (Baseline 3)	4	**92.85**	**92.10**	**91.63**	**91.08**
Config 4	5	92.04	91.88	91.32	90.75
Config 5	6	91.56	91.72	91.07	90.31

4.4. Optimizing number of iterations

Here and below, we applied the configurations that worked best for the validation set of the RUSAVIC corpus as the new baseline.

In Table 6, we analyzed the effect of varying the number of refinement process iterations (N), to identify the optimal number of refinement steps necessary to unleash the model's potential.

Table 6 shows that increasing the number of iterations enhanced the model's recognition accuracy on both the RUSAVIC and LRW corpora. This resulted in the attainment of maximal accuracy of 92.85% (Val) and 92.10% (Test) for RUSAVIC and 91.63% (Val) and 91.08% (Test) for LRW. Further iterations did not lead to substantial accuracy improvements consistently. The best accuracy for the Test set was achieved in "Config 3" using four N iterations; this emphasizes the significance of determining the suitable number of refinement iterations to enhance the recognition capabilities of AVCRFormer.

4.5. Optimizing gated mechanism

As the next step of fine-tuning the model, we considered the influence of the gated mechanism hyperparameter (α) on model performance. The α hyperparameter determines the weight distribution between the original and the adapted prototypes during both training and inference. The detailed results of recognition accuracy under different gated mechanism configurations are presented in Table 7.

The analysis of the gated mechanism hyperparameter (α) on the performance of the AVCRFormer method revealed unique patterns. Modifying the value of α led to dissimilar recognition accuracy for the RUSAVIC and LRW corpora, as presented in Table 7.

The "Config 4" provided optimal results for RUSAVIC with the hyperparameter α set to 0.5, resulting in a recognition accuracy of 94.30% (Val) and 92.90% (Test). Similarly, for the LRW corpus, when "Config 2" had the α value of 0.3, it achieved the highest recognition accuracy at 92.11% (Val) and 92.40% (Test). It is worth noting that any subsequent increase or decrease of α reduced the accuracy, which highlights a significant role of the gated mechanism hyperparameter in AVCRFormer performance.

Table 7
Impact of gated mechanism (α) on recognition accuracy (%) on the RUSAVIC and LRW corpora.

Experiment	Gated mechanism (α)	RUSAVIC		LRW	
		Val	Test	Val	Test
Baseline 3 (Default)	0.1	92.85	92.10	91.63	91.08
Config 1	0.2	93.50	92.31	92.49	91.56
Config 2	0.3	93.87	92.53	92.11	**92.40**
Config 3	0.4	94.14	92.63	**92.57**	92.26
Config 4	0.5	94.30	**92.90**	91.61	91.33
Config 5 (Baseline 4)	0.6	**94.57**	92.74	91.18	91.05
Config 6	0.7	94.19	92.47	91.12	90.91
Config 7	0.8	93.60	92.37	90.89	90.70
Config 8	0.9	92.80	92.04	90.48	90.42

Table 8
Impact of optimizers and LRs on recognition accuracy (%) on the RUSAVIC and LRW corpora.

Experiment	Optimizer	LR	RUSAVIC		LRW	
			Val	Test	Val	Test
Baseline 4 (Default)		1×10^{-3}	94.57	92.74	91.18	91.05
Config 1	Adam	1×10^{-4}	95.65	94.14	93.37	93.16
Config 2		1×10^{-5}	95.22	93.98	93.04	92.97
Config 3		1×10^{-3}	94.73	93.71	92.88	93.06
Config 4 (Baseline 5)	AdamW	1×10^{-4}	**96.02**	**96.18**	93.80	93.54
Config 5		1×10^{-5}	95.54	95.75	93.29	93.08
Config 6		1×10^{-3}	94.46	93.23	91.84	91.72
Config 7	SGD	1×10^{-4}	81.67	76.40	70.46	69.99
Config 8		1×10^{-5}	67.42	61.18	49.38	48.57
Config 9		1×10^{-3}	95.05	94.14	93.74	93.85
Config 10	Lion	1×10^{-4}	95.86	95.48	**96.09**	**96.60**
Config 11		1×10^{-5}	95.65	95.27	94.61	94.64

4.6. Optimizing training parameters

After optimizing the model architecture, refinement iterations, and gated mechanism, we focused on fine-tuning the training parameters. This stage is aimed at determining the most effective combination of LRs, optimizers, and LR schedulers to improve the convergence speed of the model and overall AVCRFormer performance.

4.6.1. Learning rates and optimizers

We compared popular optimizers: Adam (Kingma & Ba, 2014), AdamW (Loshchilov & Hutter, 2019), SGD (Ruder, 2016), and Lion (Chen, Liang, et al., 2023). We performed a systematic evaluation of three LRs for each optimizer: 1×10^{-3}, 1×10^{-4}, and 1×10^{-5}.

The results shown in Table 8 provide valuable insights into the impact of different optimizers and LRs on the recognition accuracy within both the RUSAVIC and LRW corpora. The configurations using the AdamW and Lion optimizers with 1×10^{-4} LR showed superior performance.

4.6.2. Learning rate schedulers

With the selected optimizer from the previous step, we proceeded to assess various LR schedulers presented in Table 9. We considered the following schedulers (https://pytorch.org/docs/stable/optim.html accessed on 1 Dec. 2023): StepLR, ExponentialLR, ReduceLROnPlateau, and CosineAnnealingWarmRestarts. Each scheduler was fine-tuned with different hyperparameter values.

Therefore, our single-prediction strategy produced noteworthy recognition accuracy rates of 97.63% (Val) and 98.23% (Test) for the RUSAVIC corpus, and 96.09% (Val) and 96.60% (Test) for the LRW corpus. We presented a thorough comparative analysis of our method, which involves both single- and multi-prediction strategies, compared to the SOTA methods.

4.7. Multi-prediction strategy vs. Single-prediction strategy

The comparison of single- and multi-prediction strategies with the SOTA results is shown in Table 10.

Table 10 shows that for both corpora, the multi-prediction strategy outperforms the single-prediction strategy by 0.64% to 2.21%, depending on the corpus used. We also compare the performance of the unimodal methods by setting the features of a modality to zero with the unimodal SOTA methods. For the RUSAVIC corpus, the audio- and visual-based methods using the multi-prediction strategy outperform the methods presented in Axyonov et al. (2024). For the LRW corpus, our audio-based method is inferior to the method (Petridis et al., 2018), while the visual-based method is inferior to the method (Ma et al., 2022) using an ensemble of models. At the same time, for the RUSAVIC corpus, an accuracy of 98.87% was achieved, outperforming the SOTA method by 0.22%. For the LRW corpus, 98.81% accuracy was achieved, outperforming the SOTA method by 0.05%. The results allow concluding that increasing the computational cost of the proposed model improves the accuracy of word and utterance recognition. Moreover, several predictions for each instance of AV data at different iterations of the model result in a more robust AVCRFormer method in contrast to single-predictions obtained at the final iteration. The multi-prediction strategy can be compared to the AVSR ensemble method (Chen, Hu, et al., 2023), but unlike our method, this method uses the average of probabilities from two models.

We also report the computational cost of our models. Our models take 148 MB and have 38 million parameters. The feature extraction lasts about 16 h. Training the regulated transformer takes approximately 48 min per epoch (or about 3 days for 100 epoch) using one GeForce RTX 3090 Ti with 24 GB memory. Our method operates in a real-time mode, with a processing real-time factor of 0.11 on a GPU GeForce RTX 3090 Ti and 0.29 on a CPU Intel i9-11900KF. At the same time, 1/3 of the work of the method involves processing audio signals. Times are provided for LRW in comparison with the SOTA methods. Other researchers did not include all the computational costs we mentioned; however, we can note that a more powerful GPU or more than one GPU were used to train the models. For example, the authors (Pan et al., 2022) used four NVIDIA A100 GPUs with a total of 160 GB memory to train their models during a week. Thus, we accelerated the model training process by pre-extracting features, and only one GPU was needed to train our model, unlike other methods (Kim et al., 2022; Pan et al., 2022; Zhang, Yang, et al., 2020).

4.8. Attention visualization approach

We developed the following approach to transformer attention visualization to show how the different modalities (in our case audio and video) affect SR results. The maximum results of the performance measure on both corpora are achieved when five encoders and decoders are used. Each encoder and decoder has its own feature representation for the input signal and probability prediction based on it. To evaluate which features (audio or video) are more informative to determine the predicted class in each individual encoder, we constructed feature heatmaps for the three samples from the RUSAVIC corpus. The results for the LRW corpus exhibited similar trends; hence, we concentrate our analysis on three specific samples from the RUSAVIC corpus.

As mentioned above general matrix $\boldsymbol{O} \in \mathbb{R}^{2T \times C}$ represents the combined AV feature matrix, where $2T$ are the timesteps or the number of feature vectors, and C is the number of acoustic and visual features. To create heatmaps, we again divide the $\boldsymbol{O}$ matrix into two feature matrices (one for audio features, the other for video features). This division helps illustrate the importance of features within each individual matrix more effectively. The resulting heatmaps are shown in Fig. 3.

The result visualization shows that the information content of the features in five encoders differs depending on the input signal. In the first sample, the model pays more attention to audio features

Table 9
Impact of LR schedulers on recognition accuracy (%) on the RUSAVIC and LRW corpora.

Experiment	LR Scheduler	Accuracy (%)			
		RUSAVIC		LRW	
		Val	Test	Val	Test
Baseline 5 (Default)	Constant	96.02	96.18	93.80	93.54
Config 1 (Best)	StepLR ($step_{size} = 20, gamma = 0.1$)	**97.63**	**98.23**	**96.09**	**96.60**
Config 2	ExponentialLR ($gamma = 0.1$)	96.18	96.56	95.53	95.84
Config 3	ReduceLROnPlateau ($factor = 0.1$)	95.59	96.08	94.62	95.01
Config 4	CosineAnnealingWarmRestarts ($T_0 = 10, eta_{min} = 1 \times 10^{-5}$)	96.83	97.26	95.51	96.03

Table 10
Comparison of performance (accuracy, %) of single-prediction strategy (SP) and multi-prediction strategy (MP) with the SOTA methods. Results are reported for the RUSAVIC and LRW corpora. A, V, AV refer to audio, video and audio–visual modalities.

Method	RUSAVIC, V	LRW, V
Pan et al. (2022)	–	85.00
Zhang, Yang, Xiao, Shan, and Chen (2020)	–	85.02
Martinez, Ma, Petridis, and Pantic (2020)	–	85.30
Kim, Hong, Park, and Ro (2021)	–	85.40
Feng, Yang, Shan, and Chen (2020)	–	88.40
Ma, Martinez, Petridis, and Pantic (2021)	–	88.50
Kim, Yeo, and Ro (2022)	–	88.50
Ivanko, Ryumin, et al. (2022)	–	88.70
Koumparoulis and Potamianos (2022)	–	89.52
Ma, Wang, Petridis, Shen, and Pantic (2022) (ensemble)	–	**94.10**
Axyonov, Ryumina, Ryumin, Ivanko, and Karpov (2023)	83.46	–
Method	**RUSAVIC, A/V/AV**	**LRW, A/V/AV**
Petridis et al. (2018)	–	**97.70**/82.00/98.00
Miao, Liu, and Yang (2020)	–	–/82.80/98.30
Ryumin et al. (2023)	–	96.07/87.16/98.76
Axyonov et al. (2024)	95.61/83.46/98.65	–
Ours (with SP)	94.89/82.31/98.23	94.45/88.36/96.60
Ours (with MP)	**95.75/84.03/98.87**	97.25/89.57/**98.81**

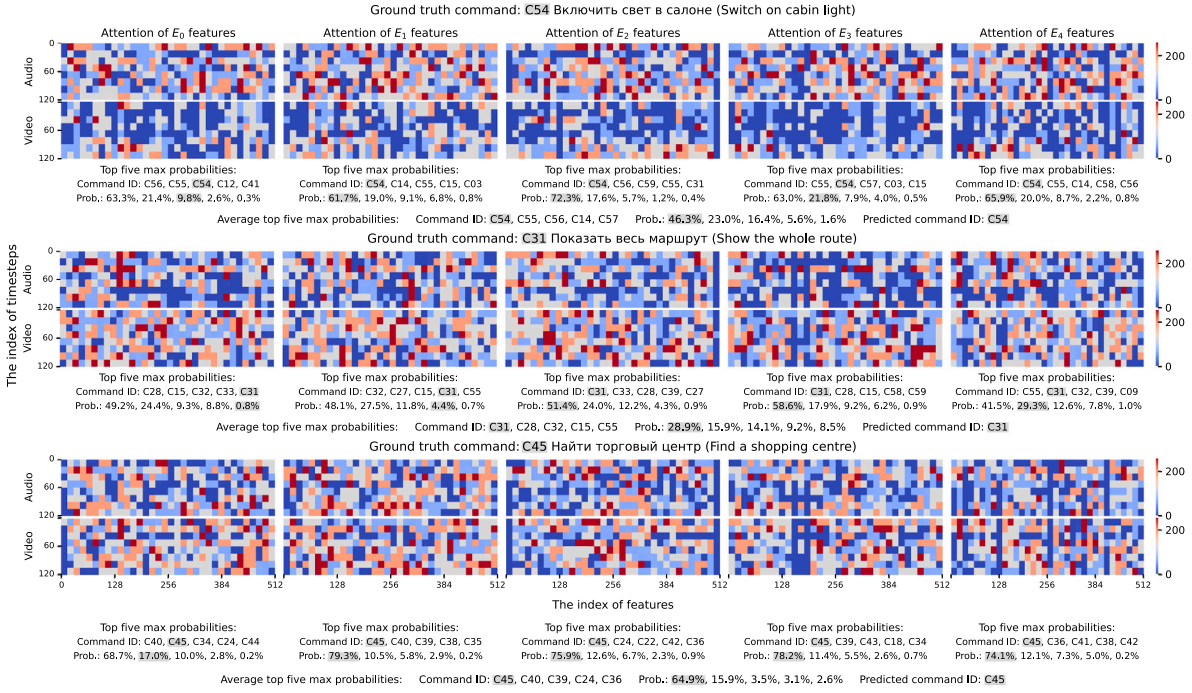


Fig. 3. Attention Visualization: heatmap examples for the RUSAVIC samples. Gray shows target classes and their predictions for each heatmap.

(more orange and bright red pixels) than to video ones, as the information content of the video modality is degraded because of lip region occlusion resulted from head rotation. In contrast, in the second sample, the model pays more attention to video features than to audio ones because of the noisy road environment, which makes the audio modality less distinguishable. Despite the loss of useful information in one of the modalities, the model still finds informative features in the weaker modalities that influence the prediction. In the cases where both

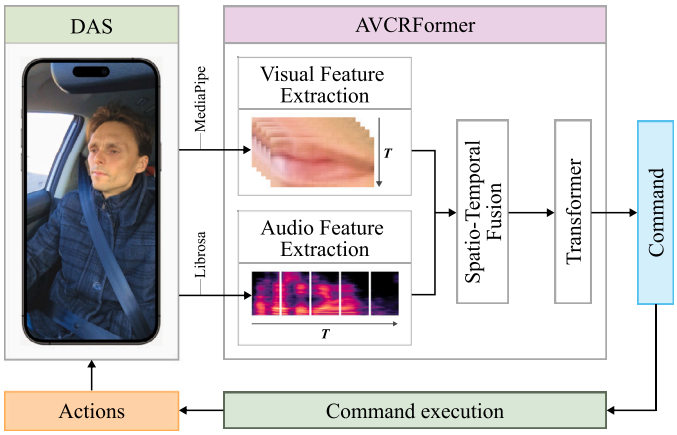


Fig. 4. AVCRFormer application for driver assistive systems (DAS).

modalities are not distorted by various factors, the model pays equal attention to the features of both modalities. Therefore, each encoder pays attention to its features regardless of the analyzed modality, so the use of multiple encoders allows the model to make predictions based on the informativeness of different features.

To visualize how predictions change iteratively based on feature importance, we also gray out target classes and their predictions in Fig. 3. Thus, for all three input samples, the $\boldsymbol{E}_0$ features of the model initialization make incorrect probability predictions, while the $\boldsymbol{E}_1$-$\boldsymbol{E}_5$ features of the iterative model refinement module result in more accurate predictions. In addition, the model based on the latest features still makes errors, but to a lesser extent. Despite the number of errors for the input samples presented, the averaging of the probability predictions leads to the correct predicted command IDs. The decoding of all command IDs in the RUSAVIC corpus is available at (https://smil-spcras.github.io/AVCRFormer/ accessed on Feb. 19, 2024).

Hence, if important information is lost in one encoder, it is taken into account in another encoder, and the prediction is based on different features. We have developed a model that is capable of: (i) compensating for the weaknesses of unimodal information, i.e., adapting to the informativeness of the modalities, (ii) making multiple predictions depending on the informativeness of the features on each encoder. The proposed multimodal architecture using multiple encoders and decoders provides stability and flexibility in processing input signals with uncontrolled recording conditions.

5. Avcrformer application for driver assistive systems

In this section, we introduce AVCRFormer-based multimodal interaction application designed to aid vehicle drivers, enhance their driving experience and ultimately increase overall road safety. This mobile application provides voice control functionalities to navigate and manage multimedia systems of smartphones, potentially extending to vehicle features such as lighting and climate control.

It is designed to enhance the accuracy and robustness of SR for common driver requests in real-world driving scenarios. Given the inherent challenges of SR in driving conditions, including ambient noise, head movements, varying poses, distance from recording devices, and lighting conditions, we leverage multimodal information and improve upon existing recognition systems, e.g. Ivanko, Ryumin, et al. (2022), by using the SOTA AVCRFormer methodology, presented in the article.

Fig. 4 illustrates the general architecture of the AVCRFormer application comprising two primary parts. The AVCRFormer part is responsible for processing results of speech and audio–visual command recognition; the actions part provides execution of voice commands and AV feedback to the driver. The software application has been developed and trained using the RUSAVIC and LRW corpora, and rigorous testing has been conducted to evaluate its functionalities and user interface.

The software application has been integrated into the Drive Safely assistive system in Ivanko, Kashevnik, et al. (2022), which provides voice control capabilities for smartphone navigation/multimedia systems. It allows possible extension to control vehicle systems, e.g. lighting conditions and climate control via voice commands. Source code, dataset and trained models are free available by request via (https://mobiledrivesafely.com, accessed on Feb. 19, 2024).

AVSR has made significant strides in improving human–computer interaction, particularly in environments like driver assistive systems where traditional unimodal speech recognition systems may falter due to acoustic noises. However, the challenge of noise in real-world applications remains a substantial hurdle:

1. **Variability and intensity of acoustic noises**: In the context of driver assistive systems, the vehicle's interior can be fraught with various noise sources, including road noise, engine sounds, conversations, and the audio output from devices.
2. **Poor lighting conditions and visual occlusions**: While visual information can supplement audio cues, especially in noisy conditions, visual noise and challenges such as poor lighting conditions or occlusions (e.g., the driver's mouth being partially covered) degrade the quality of visual data.
3. **Non-stationary noises**: The noise encountered in driving environments is predominantly non-stationary, meaning its characteristics change over time.
4. **Synchronization issues**: The integration of audio and visual data requires precise synchronization to ensure that the temporal alignment between spoken words and visual lip movements is maintained.

To effectively address the challenges of noises in real-world applications of the AVSR systems, particularly in driver assistive technologies, a multifaceted approach is essential. This involves deploying advanced noise suppression techniques to clean audio, improving spatio-temporal fusion methods for better integration of audio and visual data, developing robust visual feature extraction algorithms for poor lighting conditions and occlusions, and implementing cross-modal compensation strategies for system resilience by leveraging strong cues from one modality to compensate for weaknesses in another one.

Traditional AVSR systems address some of the noise issues described above. However, such systems often rely on separately processing audio and visual signals and then combine these at a later stage without dynamic weighting of modality importance. These systems are difficult to adapt to real-world applications because they cannot deal with different qualities of audio and video data. This occurs because the models have not been trained to adjust attention based on the informativeness of audio and video data by their simultaneous analysis.

The proposed regulated transformer model with a cross-attention mechanism dynamically adjusts the focus between audio and visual inputs based on their contextual relevance and quality. This is clearly illustrated in Fig. 3. For example, in the case of noisy environments the attention shifts to the video modality, but in the case of poor lighting conditions or high lip occlusions it shifts to the audio one. Our method outperforms the traditional ones by:

1. **Enhancing contextual understanding**: Leveraging the transformer's ability to model complex relationships within and across modalities, leading to a richer understanding of the speech context.
2. **Dynamic modality weighting**: Automatically giving precedence to clean modality (audio or visual) in real-time, thus maintaining high recognition accuracy even when one modality is compromised (e.g., noisy environments or poor visibility).

3. **Iterative model refinement**: Iteratively accounting important features missed in previous iterations and making predictions instead of solely relying on potentially inaccurate predictions from the last iteration.

Therefore, the proposed method improves upon the existing AVSR applications by introducing the regulated transformer, the fusion strategies of AV features and the classifier ensemble that account for temporal and spatial dimensions, and enhanced transparency and adaptability through attention visualization. These innovations address critical challenges in AVSR, particularly for applications in noisy and visually complex environments like those encountered in driver assistive systems, paving the way for more accurate, robust, and user-friendly SR technologies.

6. Conclusions

In this article, we proposed a novel AVSR method aimed at enhancing SR within driver assistive systems, crucial for ensuring safe interaction while driving. The proposed method introduces a novel transformer model, AVCRFormer, tailored specifically for robust AVSR. Key innovations include a multimodal fusion strategy, a regulated transformer architecture, and a classifier ensemble strategy. By preserving contextual information and leveraging the iterative refinement of models, AVCRFormer demonstrates superior performance in speech command recognition accuracy. The implementation of a multi-prediction strategy demonstrates superior adaptability in diverse AV contexts. Achieving high accuracy rates of 98.87% and 98.81% on the RUSAVIC and LRW corpora respectively, this research significantly contributes to advancing human–computer interaction and holds promise for broader applications in audio–visual processing and artificial intelligence.

Although our method has demonstrated proficiency in semi-controlled environments, its transition to real-world applications presents significant challenges, such as latency, computational efficiency, and scalability. Our future endeavors will prioritize optimizing AVCRFormer for seamless integration into practical scenarios, with a focus on addressing challenges specific to the automotive environment. In addition, we aim to deploy AVCRFormer for different acoustic environments, speech patterns, and languages.

CRediT authorship contribution statement

Dmitry Ryumin: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft, Project administration. **Alexandr Axyonov:** Conceptualization, Methodology, Visualization, Writing – original draft. **Elena Ryumina:** Conceptualization, Methodology, Investigation, Visualization, Writing – original draft. **Denis Ivanko:** Conceptualization, Resources, Writing – original draft, Writing – review & editing. **Alexey Kashevnik:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing. **Alexey Karpov:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Analytical Center for the Government of the Russian Federation (IGK 000000D730324P540002), agreement No. 70-2021-00141.

References

Adeel, A., Gogate, M., & Hussain, A. (2020). Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments. *Information Fusion*, *59*, 163–170. http://dx.doi.org/10.1016/j.inffus.2019.08.008.

Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 8717–8727. http://dx.doi.org/10.1109/TPAMI.2018.2889052.

Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: A large-scale dataset for visual speech recognition. (pp. 1–2). http://dx.doi.org/10.48550/arXiv.1809.00496, arXiv preprint arXiv:1809.00496.

Alvarez, I., Martin, A. M., Dunbar, J., Taiber, J. G., Wilson, D.-M., & Gilbert, J. E. (2010). Voice interfaced vehicle user help. In *International conference on automotive user interfaces and interactive vehicular applications* (pp. 42–49). http://dx.doi.org/10.1145/1969773.1969782.

Anwar, M., Shi, B., Goswami, V., Hsu, W.-N., Pino, J., & Wang, C. (2023). MuAViC: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. (pp. 1–8). http://dx.doi.org/10.48550/arXiv.2303.00628, ArXiv.

Axyonov, A., Ryumin, D., Ivanko, D., Kashevnik, A., & Karpov, A. (2024). Audio-visual speech recognition in-the-Wild: Multi-angle vehicle cabin corpus and attention-based method. In *IEEE international conference on acoustics, speech and signal processing* (pp. 8195–8199). IEEE, http://dx.doi.org/10.1109/ICASSP48485.2024.10448048.

Axyonov, A., Ryumina, E., Ryumin, D., Ivanko, D., & Karpov, A. (2023). Neural network-based method for visual recognition of driver's voice commands using attention mechanism. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, *23*(4), 767–775. http://dx.doi.org/10.17586/2226-1494-2023-23-4-767-775.

Berkol, A., Tümer-Sivri, T., Pervan-Akman, N., Çolak, M., & Erdem, H. (2023). Visual lip reading dataset in Turkish. *Data*, *8*(1), 15. http://dx.doi.org/10.3390/data8010015.

Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., et al. (2022). Automatic speech recognition (ASR) systems for children: A systematic literature review. *Applied Sciences*, *12*(9), 4419. http://dx.doi.org/10.3390/app12094419.

Biswas, A., Sahu, P. K., & Chandra, M. (2015). Multiple camera in car audio–visual speech recognition using phonetic and visemic information. *Computers & Electrical Engineering*, *47*, 35–50. http://dx.doi.org/10.1016/j.compeleceng.2015.08.009.

Braun, M., Mainz, A., Chadowitz, R., Pfleging, B., & Alt, F. (2019). At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). http://dx.doi.org/10.1145/3290605.3300270.

Burchi, M., & Timofte, R. (2023). Audio-visual efficient conformer for robust speech recognition. In *IEEE/CVF winter conference on applications of computer vision* (pp. 2258–2267). http://dx.doi.org/10.1109/WACV56688.2023.00229.

Chen, C., Hu, Y., Zhang, Q., Zou, H., Zhu, B., & Chng, E. S. (2023). Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning. In *AAAI conference on artificial intelligence* (pp. 12607–12615). http://dx.doi.org/10.48550/arXiv.2212.05301.

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., et al. (2023). Symbolic discovery of optimization algorithms. (pp. 1–30). http://dx.doi.org/10.48550/arXiv.2302.06675, arXiv preprint arXiv:2302.06675.

Chen, C., Wang, D., & Zheng, T. F. (2023). CN-CVS: A mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis. In *IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). IEEE, http://dx.doi.org/10.1109/ICASSP49357.2023.10095796.

Cheng, M., Wang, H., Wang, Y., & Li, M. (2022). The DKU audio-visual wake word spotting system for the 2021 MISP challenge. In *IEEE international conference on acoustics, speech and signal processing* (pp. 9256–9260). http://dx.doi.org/10.1109/icassp43922.2022.9747216.

Chuang, S.-Y., Wang, H.-M., & Tsao, Y. (2022). Improved lite audio-visual speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 1345–1359. http://dx.doi.org/10.1109/taslp.2022.3153265.

Chung, J. S., Senior, A. W., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 3444–3453). http://dx.doi.org/10.1109/CVPR.2017.367.

Chung, J. S., & Zisserman, A. (2016). Lip reading in the wild. In *Asian conference on computer vision* (pp. 87–103). http://dx.doi.org/10.1007/978-3-319-54184-6_6.

Dai, W., Cahyawijaya, S., Yu, T., Barezi, E. J., Xu, P., Yiu, C. T. S., et al. (2022). CI-AVSR: A cantonese audio-visual speech dataset for in-car command recognition. In *International conference on language resources and evaluation* (pp. 6786–6793).

Dai, Y., Chen, H., Du, J., Ding, X., Ding, N., Jiang, F., et al. (2023). Improving audio-visual speech recognition by lip-subword correlation based visual pre-training and cross-modal fusion encoder. In *IEEE international conference on multimedia and expo* (pp. 2627–2632). IEEE, http://dx.doi.org/10.1109/ICME55011.2023.00447.

Deng, K., Cao, S., Zhang, Y., & Ma, L. (2021). Improving hybrid CTC/Attention end-to-end speech recognition with pretrained acoustic and language models. In *IEEE automatic speech recognition and understanding workshop* (pp. 76–82). IEEE, http://dx.doi.org/10.1109/ASRU51503.2021.9688009.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Egorov, E., Kostyumov, V., Konyk, M., & Kolesnikov, S. (2021). LRWR: Large-scale benchmark for lip reading in Russian language. (pp. 1–5). http://dx.doi.org/10.48550/arXiv.2109.06692, arXiv preprint arXiv:2109.06692.

Feng, D., Yang, S., Shan, S., & Chen, X. (2020). Learn an effective lip reading model without pains. (pp. 1–6). http://dx.doi.org/10.48550/arXiv.2011.07557, arXiv preprint arXiv:2011.07557.

Gerczuk, M., Amiriparian, S., Ottl, S., & Schuller, B. (2021). EmoNet: A transfer learning framework for multi-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*, *14*, 1472–1487. http://dx.doi.org/10.1109/TAFFC.2021.3135152.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., et al. (2020). Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech* (pp. 5036–5040). http://dx.doi.org/10.21437/Interspeech.2020-3015.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778). http://dx.doi.org/10.1109/cvpr.2016.90.

Hong, J., Kim, M., Choi, J. Y., & Ro, Y. M. (2023). Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 18783–18794). http://dx.doi.org/10.1109/CVPR52729.2023.01801.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech rrepresentation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460. http://dx.doi.org/10.1109/TASLP.2021.3122291.

Isobe, S., Tamura, S., Hayamizu, S., Gotoh, Y., & Nose, M. (2021). Multi-angle lipreading with angle classification-based feature extraction and its application to audio-visual speech recognition. *Future Internet*, *13*(7), 182. http://dx.doi.org/10.3390/fi13070182.

Ivanko, D., Axyonov, A., Ryumin, D., Kashevnik, A., & Karpov, A. (2022). RUSAVIC corpus: Russian audio-visual speech in cars. In *Language resources and evaluation conference* (pp. 1555–1559).

Ivanko, D., Kashevnik, A., Ryumin, D., Kitenko, A., Axyonov, A., Lashkov, I., et al. (2022). MIDriveSafely: Multimodal interaction for drive safely. In *Proceedings of the 2022 international conference on multimodal interaction* (pp. 733–735). http://dx.doi.org/10.1145/3536221.3557037.

Ivanko, D., Ryumin, D., Kashevnik, A., Axyonov, A., & Karpov, A. (2022). Visual speech recognition in a driver assistance system. In *European signal processing conference* (pp. 1131–1135). IEEE, http://dx.doi.org/10.23919/eusipco55093.2022.9909819.

Ivanko, D., Ryumin, D., Kashevnik, A., Axyonov, A., Kitenko, A., Lashkov, I., et al. (2022). DAVIS: Driver's audio-visual speech recognition. In *ISCA annual conference interspeech* (pp. 1141–1142).

Jaradat, G. A., Alzubaidi, M. A., & Otoom, M. (2022). A novel human-vehicle interaction assistive device for Arab drivers using speech recognition. *IEEE Access*, *10*, 127514–127529. http://dx.doi.org/10.1109/ACCESS.2022.3226539.

Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., et al. (2020). Libri-light: A benchmark for asr with limited or no supervision. In *Proceedings of the 2020 international conference on acoustics, speech and signal processing* (pp. 7669–7673). IEEE.

Katsaggelos, A. K., Bahaadini, S., & Molina, R. (2015). Audiovisual fusion: Challenges and new approaches. *IEEE*, *103*, 1635–1653. http://dx.doi.org/10.1109/JPROC.2015.2459017.

Khan, S. H., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, *54*(10s), 1–41. http://dx.doi.org/10.1145/3505244.

Kim, M., Hong, J., Park, S. J., & Ro, Y. M. (2021). Multi-modality associative bridging through memory: Speech sound recollected from face video. In *IEEE/CVF international conference on computer vision* (pp. 296–306). http://dx.doi.org/10.1109/ICCV48922.2021.00036.

Kim, S., Hori, T., & Watanabe, S. (2017). Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *IEEE international conference on acoustics, speech and signal processing* (pp. 4835–4839). IEEE, http://dx.doi.org/10.1109/ICASSP.2017.7953075.

Kim, M., Yeo, J. H., & Ro, Y. M. (2022). Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *AAAI conference on artificial intelligence* (pp. 1174–1182). http://dx.doi.org/10.48550/arXiv.2204.01725.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. (pp. 1–15). http://dx.doi.org/10.48550/arXiv.1412.6980, arXiv preprint arXiv:1412.6980.

Koumparoulis, A., & Potamianos, G. (2022). Accurate and resource-efficient lipreading with Efficientnetv2 and transformers. In *IEEE international conference on acoustics, speech and signal processing* (pp. 8467–8471). IEEE, http://dx.doi.org/10.1109/ICASSP43922.2022.9747729.

Kumar, L. A., Renuka, D. K., Rose, S. L., Wartana, I. M., et al. (2022). Deep learning based assistive technology on audio visual speech recognition for hearing impaired. *International Journal of Cognitive Computing in Engineering*, *3*, 24–30. http://dx.doi.org/10.1016/j.ijcce.2022.01.003.

Lee, Y.-H., Jang, D.-W., Kim, J.-B., Park, R.-H., & Park, H.-M. (2020). Audio-visual speech recognition based on dual cross-modality attentions with the transformer model. *Applied Sciences*, *10*(20), 7263. http://dx.doi.org/10.3390/app10207263.

Li, J., Li, C., Wu, Y., & Qian, Y. (2023). Robust audio-visual ASR with unified cross-modal attention. In *IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). http://dx.doi.org/10.1109/ICASSP49357.2023.10096893.

Li, T., Zhou, H., Wang, J., Hong, Q., & Li, L. (2023). The XMU system for audio-visual diarization and recognition in MISP challenge 2022. In *IEEE international conference on acoustics, speech and signal processing* (pp. 1–2). http://dx.doi.org/10.1109/ICASSP49357.2023.10095693.

Li, J., et al. (2022). Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, *11*(1), http://dx.doi.org/10.1561/116.00000050.

Liang, B., & Tang, Z. (2022). Multi-modal information analysis of automobile intelligent human-computer interaction. In *International conference on multi-modal information analytics* (pp. 658–666). Springer, http://dx.doi.org/10.1007/978-3-031-05237-8_81.

Lin, Y.-B., Sung, Y.-L., Lei, J., Bansal, M., & Bertasius, G. (2023). Vision transformers are parameter-efficient audio-visual learners. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 2299–2309). IEEE, http://dx.doi.org/10.1109/CVPR52729.2023.00228.

Liu, H., Li, W., & Yang, B. (2021). Robust audio-visual speech recognition based on hybrid fusion. In *International conference on pattern recognition* (pp. 7580–7586). IEEE, http://dx.doi.org/10.1109/ICPR48806.2021.9412817.

Loh, C. Y., Boey, K. L., & Hong, K. S. (2017). Speech recognition interactive system for vehicle. In *2017 IEEE 13th international colloquium on signal processing & its applications* (pp. 85–88). http://dx.doi.org/10.1109/CSPA.2017.8064929.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations* (pp. 1–18).

Lu, Y., Li, Z., He, D., Sun, Z., Dong, B., Qin, T., et al. (2019). Understanding and improving transformer from a multi-particle dynamic system point of view. (pp. 1–15). http://dx.doi.org/10.48550/arXiv.1906.02762, arXiv preprint arXiv:1906.02762.

Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., & Pantic, M. (2023). Auto-AVSR: Audio-visual speech recognition with automatic labels. In *IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). IEEE, http://dx.doi.org/10.1109/ICASSP49357.2023.10096889.

Ma, P., Martinez, B., Petridis, S., & Pantic, M. (2021). Towards practical lipreading with distilled and efficient models. In *IEEE international conference on acoustics, speech and signal processing* (pp. 7608–7612). IEEE, http://dx.doi.org/10.1109/ICASSP39728.2021.9415063.

Ma, P., Petridis, S., & Pantic, M. (2021). End-to-end audio-visual speech recognition with conformers. In *IEEE international conference on acoustics, speech and signal processing* (pp. 7613–7617). IEEE, http://dx.doi.org/10.1109/ICASSP39728.2021.9414567.

Ma, P., Wang, Y., Petridis, S., Shen, J., & Pantic, M. (2022). Training strategies for improved lip-reading. In *IEEE international conference on acoustics, speech and signal processing* (pp. 8472–8476). IEEE, http://dx.doi.org/10.1109/ICASSP43922.2022.9746706.

Makino, T., Liao, H., Assael, Y., Shillingford, B., Garcia, B., Braga, O., et al. (2019). Recurrent neural network transducer for audio-visual speech recognition. In *IEEE automatic speech recognition and understanding workshop* (pp. 905–912). IEEE, http://dx.doi.org/10.1109/ASRU46091.2019.9004036.

Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020). Lipreading using temporal convolutional networks. In *IEEE international conference on acoustics, speech and signal processing* (pp. 6319–6323). IEEE, http://dx.doi.org/10.1109/ICASSP40776.2020.9053841.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., et al. (2015). Librosa: Audio and music signal analysis in Python. In *Python in science conference*: *vol. 8*, (pp. 18–25). http://dx.doi.org/10.25080/MAJORA-7B98E3ED-003.

Mehler, B., Kidd, D. G., Reimer, B., Reagan, I. J., Dobres, J., & McCartt, A. T. (2015). Multi-modal assessment of on-road demand of voice and manual phone calling and voice navigation entry across two embedded vehicle systems. *Ergonomics*, *59*, 344–367. http://dx.doi.org/10.1080/00140139.2015.1081412.

Miao, H., Cheng, G., Gao, C., Zhang, P., & Yan, Y. (2020). Transformer-based online CTC/Attention end-to-end speech recognition architecture. In *IEEE international conference on acoustics, speech and signal processing* (pp. 6084–6088). IEEE, http://dx.doi.org/10.1109/ICASSP40776.2020.9053165.

Miao, Z., Liu, H., & Yang, B. (2020). Part-based lipreading for audio-visual speech recognition. In *IEEE international conference on systems, man, and cybernetics* (pp. 2722–2726). IEEE, http://dx.doi.org/10.1109/smc42975.2020.9283044.

Michelsanti, D., Tan, Z.-H., Zhang, S.-X., Xu, Y., Yu, M., Yu, D., et al. (2021). An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1368–1396. http://dx.doi.org/10.1109/TASLP.2021.3066303.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *42*, 722–737. http://dx.doi.org/10.1007/s10489-014-0629-7.

Ouyang, R., Ma, C., & Tan, H. (2023). The development of China automotive human-computer interaction. In *International conference on human-computer interaction* (pp. 417–429). Springer, http://dx.doi.org/10.1007/978-3-031-35939-2_31.

Pan, X., Chen, P., Gong, Y., Zhou, H., Wang, X., & Lin, Z. (2022). Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. In *Annual meeting of the association for computational linguistics* (pp. 4491–4503). http://dx.doi.org/10.18653/v1/2022.acl-long.308.

Park, J. G., Hwang, J.-W., Choi, K., Lee, S.-H., Ahn, J.-H., Park, R.-H., et al. (2023). OLKAVS: An open large-scale Korean audio-visual speech dataset. (pp. 1–15). http://dx.doi.org/10.48550/arXiv.2301.06375, ArXiv arXiv:2301.06375.

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018). End-to-end audiovisual speech recognition. In *IEEE international conference on acoustics, speech and signal processing* (pp. 6548–6552). IEEE, http://dx.doi.org/10.1109/ICASSP.2018.8461326.

Peymanfard, J., Heydarian, S., Lashini, A., Zeinali, H., Mohammadi, M. R., & Mozayani, N. (2023). A multi-purpose audio-visual corpus for multi-modal Persian speech recognition: The Arman-AV dataset. *Expert Systems with Applications*, *238*, Article 121648. http://dx.doi.org/10.1016/j.eswa.2023.121648.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. (pp. 1–14). http://dx.doi.org/10.48550/arXiv.1609.04747, arXiv preprint arXiv:1609.04747.

Ryumin, D., Ivanko, D., & Ryumina, E. (2023). Audio-visual speech and gesture recognition by sensors of mobile devices. *Sensors*, *23*(4), 2284. http://dx.doi.org/10.3390/s23042284.

Ryumina, E., Dresvyanskiy, D., & Karpov, A. (2022). In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, *514*, 435–450. http://dx.doi.org/10.1016/j.neucom.2022.10.013.

Shashidhar, R., Patilkulkarni, S., & Puneeth, S. (2022). Combining audio and visual speech recognition using LSTM and deep convolutional neural network. *International Journal of Information Technology*, *14*(7), 3425–3436. http://dx.doi.org/10.1007/s41870-022-00907-y.

Shewalkar, A., Nyavanandi, D., & Ludwig, S. A. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, *9*(4), 235–245. http://dx.doi.org/10.2478/jaiscr-2019-0006.

Shi, B., Hsu, W.-N., & Mohamed, A. (2022). Robust self-supervised audio-visual speech recognition. In *Interspeech* (pp. 2118–2122). http://dx.doi.org/10.21437/Interspeech.2022-99.

Shillingford, B., Assael, Y., Hoffman, M. W., Paine, T., Hughes, C., Prabhu, U., et al. (2019). Large-scale visual speech recognition. In *Interspeech* (pp. 4135–4139). http://dx.doi.org/10.21437/Interspeech.2019-1669.

Song, Q., Sun, B., & Li, S. (2022). Multimodal sparse transformer network for audio-visual speech recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 10028–10038. http://dx.doi.org/10.1109/TNNLS.2022.3163771.

Sterpu, G., Saam, C., & Harte, N. (2018). Attention-based audio-visual fusion for robust automatic speech recognition. In *20th ACM international conference on multimodal interaction* (pp. 1–6). http://dx.doi.org/10.1145/3242969.3243014.

Tamazin, M., Gouda, A., & Khedr, M. (2019). Enhanced automatic speech recognition system based on enhancing power-normalized cepstral coefficients. *Applied Sciences*, *9*(10), 2166. http://dx.doi.org/10.3390/APP9102166.

Thomas, B., Kessler, S., & Karout, S. (2022). Efficient adapter transfer of self-supervised speech models for automatic speech recognition. In *IEEE international conference on acoustics, speech and signal processing* (pp. 7102–7106). IEEE, http://dx.doi.org/10.1109/ICASSP43922.2022.9746223.

Wang, J., Huo, Y., Liu, L., Xu, T., Li, Q., & Li, S. (2023). MAVD: The First Open Large-Scale Mandarin Audio-Visual Dataset with Depth Information. In *Interspeech* (pp. 2113–2117). http://dx.doi.org/10.21437/Interspeech.2023-823.

Wang, H., Zha, Z.-J., Li, L., Chen, X., & Luo, J. (2023). Context-aware proposal-boundary network with structural consistency for audiovisual event localization. *IEEE Transactions on Neural Networks and Learning Systems*, 1–11. http://dx.doi.org/10.1109/TNNLS.2023.3290083.

Wu, Y., Li, C., Yang, S., Wu, Z., & Qian, Y. (2021). Audio-visual multi-talker speech recognition in a cocktail party. In *Interspeech* (pp. 1–5). http://dx.doi.org/10.21437/interspeech.2021-2128.

Yang, W., Li, P., Yang, W., Liu, Y., He, Y., Petrosian, O., et al. (2023). Research on robust audio-visual speech recognition algorithms. *Mathematics*, *11*(7), 1733. http://dx.doi.org/10.3390/math11071733.

Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., et al. (2019). LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *IEEE international conference on automatic face & gesture recognition* (pp. 1–8). IEEE, http://dx.doi.org/10.1109/FG.2019.8756582.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., et al. (2023). Google USM: Scaling automatic speech recognition beyond 100 languages. (pp. 1–20). http://dx.doi.org/10.48550/arXiv.2303.01037, arXiv preprint arXiv:2303.01037.

Zhang, Z., Pinto, J., Plahl, C., Schuller, B., & Willett, D. (2014). Channel mapping using bidirectional long short-term memory for dereverberation in hands-free voice controlled devices. *IEEE Transactions on Consumer Electronics*, *60*, 525–533. http://dx.doi.org/10.1109/TCE.2014.6937339.

Zhang, Y., Yang, S., Xiao, J., Shan, S., & Chen, X. (2020). Can we read speech beyond the lips? Rethinking ROI selection for deep visual speech recognition. In *IEEE international conference on automatic face and gesture recognition* (pp. 356–363). IEEE, http://dx.doi.org/10.1109/FG47880.2020.00134.

Zhang, L., Zhao, Z., Ma, C., Shan, L., Sun, H., Jiang, L., et al. (2020). End-to-end automatic pronunciation error detection based on improved hybrid CTC/Attention architecture. *Sensors*, *20*(7), 1809. http://dx.doi.org/10.3390/s20071809.

Zhou, P., Yang, W., Chen, W., Wang, Y., & Jia, J. (2019). Modality attention for end-to-end audio-visual speech recognition. In *IEEE international conference on acoustics, speech and signal processing* (pp. 6565–6569). IEEE, http://dx.doi.org/10.1109/ICASSP.2019.8683733.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2020). A comprehensive survey on transfer learning. *IEEE*, *109*(1), 43–76. http://dx.doi.org/10.1109/JPROC.2020.3004555.