Interspeech 2022
18-22 September 2022, Incheon, Korea

DAVIS: Driver's Audio-Visual Speech Recognition

Denis Ivanko[1], *Dmitry Ryumin*[1], *Alexey Kashevnik*[2], *Alexandr Axyonov*[1], *Andrey Kitenko*[1], *Igor Lashkov*[1], *Alexey Karpov*[1]

[1]St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, Russia
[2]ITMO University, Saint Petersburg, Russia
{ivanko.d, ryumin.d, alexey.kashevnik, axyonov.a, andrey.kitenko, igor.lashkov, karpov}@iias.spb.su

Abstract

DAVIS is a driver's audio-visual assistive system intended to improve accuracy and robustness of speech recognition of the most frequent drivers' requests in natural driving conditions. Since speech recognition in driving condition is highly challenging due to acoustic noises, active head turns, pose variation, distance to recording devices, lightning conditions, etc. We rely on multimodal information and use both automatic lip-reading system for visual stream and ASR for audio stream processing. We have trained audio and video models on own RUSAVIC dataset containing in-the-wild audio and video recordings of 20 drivers. The recognition application comprises a graphical user interface and modules for audio and video signal acquisition, analysis, and recognition. The obtained results demonstrate rather high performance of DAVIS and also the fundamental possibility of recognizing speech commands by using video modality, even in such difficult natural conditions as driving.

Index Terms: audio-visual speech recognition, driver assistance system, human-computer interaction

1. Introduction

Using visual information about speech in addition to audio is a fundamental step to human-like robust speech recognition system for challenging acoustic conditions [1]. Furthermore, visual data itself often contains enough information to recognize spoken phrases [2]. In the recent years, to improve recognition results some researchers employ visual data and investigate how lip-reading can contribute to audio-based speech recognition [3].

However, at the moment there is no reliable noise-robust speech recognition system to be used in real driving conditions. The use of hands to control navigation system/air conditioner/smartphone distracts a driver and cause road accidents. The acoustic noise itself is not only challenge in the domain [4]. A background noise affects not only the microphone, but also it causes the speaker to increase vocal effort to overcome noise levels in his ears (the so-called Lombard effect). In the real-world scenarios variation of speech production caused by noise exposure at the ear can damage the performance more than the acoustic noise itself [5]. In current research to train our models we use own RUSAVIC corpus that contains audio-visual speech data of 20 different drivers recorded in-the-wild [6].

In this work, we introduce DAVIS: a speech recognition assistive system for drivers that is able to recognize most frequent drivers' control commands (62) by processing acoustic and visual speech information. We present the recognition app that provides a graphical user interface and modules for audio-visual signals acquisition, analysis, and recognition. DAVIS has been experimented and evaluated in some real-world scenarios. We make source code, dataset and trained models open access.

2. DAVIS Architecture

DAVIS has been developed as a driver's assistant application intended to improve accuracy and robustness of speech recognition in challenging acoustic conditions by processing audio-visual information.

It is developed for drivers to tackle 62 the most frequent requests to navigation/multimedia systems of the vehicle [7]. We designed the system to be used inside a vehicle cabin and offer drivers the possibility of robust speech recognition despite acoustic noises, active head turns, pose variation, distance to recording devices, lightning conditions, etc.

The app has two panels: (i) setting panel, where a user can select one of speech recognition modes (audio, visual or audio-visual), select desirable recognition mode, dictionary, etc. and (ii) recognition panel (Figure 1, right) that shows camera view, detected driver's mouth region and the last recognized phrase at the overlay. Screenshot of the DAVIS setup is shown in the Figure 1, left. Currently DAVIS uses audio-visual recognition architecture and the dictionary of 62 most frequent drivers' requests.

The speech recognition pipeline of the DAVIS is shown in Figure 2. The app is implemented as a GUI integrated with five software modules: (1) audio-visual signal acquisition module, (2) voice activity detection module, (3) audio-visual features extraction module, (4) recognition models, (5) modalities fusion and recognition module.

Figure 1: *DAVIS setup (left) and app screenshot (right) during recognition process*

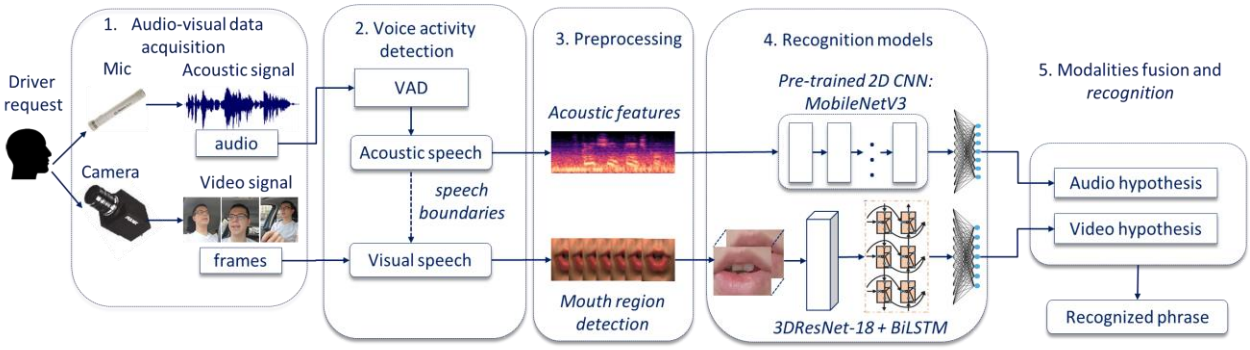

Figure 2: *DAVIS audio-visual speech recognition pipeline*

The signal acquisition module is used for capturing audio (mp4→wav) and video (mp4→ frames) signals. The audio signal is recorded at sampling frequency of 16kHz. The video of the speaker is simultaneously recorded at 30 frames per second with the resolution of 1280×720 pixels. The voice activity detection module is based on the Vosk model [8] gets acoustic speech boundaries and uses them to extract speech utterances from the raw audio/video signals.

Feature extraction (preprocessing) module performs the first step by detecting and cropping mouth images on each frame of the video, followed by some visual data preprocessing procedures: grayscaling, normalization, and histogram alignment. It also exctacts spectrograms from acoustic signal.

We use end-to-end neural network architectures for audio and visual speech recognition models. The core of the lip-reading model includes a modified 3DResNet-18 neural network [9] in followed by 2 layers of BiLSTM [10]. The core of the acoustic speech recognition is similar to the work [11].

Information fusion module performs a comparative assessment of recognition hypothesis provided by lip-reading and ASR models and makes the final decision. In our case, one of the 62 voice commands. Currently DAVIS uses weighting algorithm to this end [12]. In a simplified form: the more acoustically noisy the environment, the more we rely on video information and vice versa. If the angle of the face in relation to the camera is large or the video is dark / overexposed, then we rely more on the acoustic system. At the moment, the algorithm considers a large set of parameters, such as vehicle speed, head angle, SNR level, presence of music, etc.

3. Conclusions

Accurate speech recognition for drivers is challenging, in particular due to acoustic noise, active head turns, pose variation, distance to recording devices, lightning conditions, etc. DAVIS is an application that helps drivers by improving speech recognition accuracy of most frequent requests addressed to navigation/multimedia systems of the vehicle. The DAVIS implements both: automatic lip-reading system and acoustic speech recognition system. It has been developed as an app and trained on the real-world data. It has been tested for its functionalities and user interface. Source code, dataset and trained models free available by request[1].

For future work we plan to integrate the developed DAVIS system to our developed Drive Safely system[2] for voice-based command support [13]. Drive Safely is a driver monitoring system that detects dangerous driver behavior while driving (drowsiness, distraction, smartphone usage, and etc.).

4. Acknowledgements

This work was supported by the Analytical Center for the Government of the Russian Federation (IGK 000000D730321P5Q0002), agreement No. 70-2021-00141.

5. References

[1] B. Shi et al., “Robust Self-Supervised Audio-Visual Speech Recognition”, arXiv preprint arXiv:2201. 01763, 2022.

[2] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition”, IEEE transactions on pattern analysis and machine intelligence, 2018, pp. 1-13.

[3] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia, “Modality attention for end-to-end audio-visual speech recognition”, in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6565–6569.

[4] S.-C. Lin et al., “Adasa: A conversational in-vehicle digital assistant for advanced driver assistance features”, in Proceedings of the 31st ACM UIST, 2018, pp. 531–542.

[5] B. Lee et al., “AVICAR: Audio-visual speech corpus in a car environment”, in Eighth International Conference on Spoken Language Processing, 2004.

[6] A. Kashevnik *et al.*, “Multimodal corpus design for audio-visual speech recognition in vehicle cabin”, IEEE Access, vol 9, 2021, pp. 34986–35003.

[7] D. Ivanko et al., “Multi-Speaker Audio-Visual Corpus RUSAVIC: Russian Audio-Visual Speech in Cars”, in LREC 2022 Conference, pp. 1555-1559.

[8] Vosk Speech Recognition Toolkit. Available at: https://github.com/alphacep/vosk-api

[9] M. Kim, J. Hong, S. J. Park and Y. M. Ro, “Multi-modality associative bridging through memory: Speech sound recollected from face video”, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 296–306.

[10] A. Howard et al., “Searching for mobilenetv3”, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314-1324.

[11] A. Kashevnik, I. Lashkov, A. Gurtov “Methodology and Mobile Application for Driver Behavior Analysis and Accident Prevention”, in IEEE Transactions on Intelligent Transportation Systems, IEEE, Vol. 21(6), 2019, pp. 2427–2436.

[12] D. Ivanko et al. “Multimodal speech recognition: increasing accuracy using high speed video data”, in Journal on Multimodal User Interfaces, 12(4), 2018, pp. 319-328.

[13] D. Ivanko et al. “Visual speech recognition in a driver assistance system” in 30th European Signal Processing Conference (EUSIPCO), 2022, Accepted to EUSIPCO 2022.

[1] https://mobiledrivesafely.com

[2] https://play.google.com/store/apps/details?id=ru.igla.drivesafely