Near-infrared Lipreading System for Driver-Car Interaction

Samar Daou[1], Ahmed Rekik[1,2], Achraf Ben-Hamadou[1,2] and Abdelaziz Kallel[1,2]
[1]*Laboratory of Signals, systeMs, aRtificial Intelligence and neTworkS, Technopark of Sfax, Sakiet Ezzit, 3021 Sfax, Tunisia*
[2]*Digital Research Centre of Sfax, Technopark of Sfax, Sakiet Ezzit, 3021 Sfax, Tunisia*

Keywords: Lipreading, Audiovisual Dataset, Human-Machine Interaction, Graph Neural Networks.

Abstract: In this paper, we propose a new lipreading approach for driver-car interaction in a cockpit monitoring environment. Furthermore, we introduce and release the first lipreading dataset dedicated to intuitive driver-car interaction using near-infrared driver monitoring cameras. In this paper, we propose a two-stream deep learning architecture that combines both geometric and global visual features extracted from the mouth region to improve the performance of lipreading based only on visual cues. Geometric features are extracted by a graph convolutional network applied to a series of 2D facial landmarks, while a 2D-3D convolutional network is used to extract the global visual features from the near-infrared frame sequence. These features are then decoded based on a multi-scale temporal convolutional network to generate the output word sequence classification. Our proposed model achieved high accuracy for both training scenarios overlapped speaker and unseen speaker with 98.5% and 92.2% respectively.

1 INTRODUCTION

Lipreading, or visual speech recognition, is a method of identifying speech in a video by observing the movements of the lips and the surrounding region using only visual information. It is an impressive skill that can be used in a variety of situations, such as helping people with listening disabilities, in criminal conversations, and enhancing the performance of speech recognition systems.

Learning to read lips is challenging due basically to homophones between distinct characters (such as 'p' and 'b' in English) that produce very similar and confusing lip movement sequences at words level. Furthermore, lipreading suffers from other known challenges associated with subject dependencies like facial appearance variations and various speaking accents, speed, and manner.

Thanks to the significant advances in deep learning techniques, the field of lipreading has gained a lot of attention these last years yielding many applications (Sheng et al., 2022). Human-machine interaction is one of the prominent applications since lipreading can be used to dictate messages or instructions, especially in noisy environments or with multiple speakers. However, only a few lipreading systems for mobile device interaction have been developed (Rekik et al., 2016; Rekik et al., 2015b; Rekik et al., 2015a; Sun et al., 2018). In this research study, we are interested in designing a novel car-driver interaction system based on lipreading. This solution can be associated with voice recognition systems (Afouras et al., 2018; Ben-Hamadou, 2020) to improve their performance, especially in a noisy car environment. The majority of existing lipreading datasets are useless in this context since they were recorded with RGB cameras, while in car cockpits, infrared cameras are typically mounted to serve in all lighting conditions, both day and night.

In this paper, we present a novel lipreading-based Human-Machine interaction system for the vehicle cockpit context. In addition, we release the first publicly available lip-reading dataset dedicated to driver-car interaction, obtained with a real driver monitoring camera.

The remaining of this paper is organized as follows. Section 2 discusses related work. Our based-lipreading Human-Machine interaction system is detailed in Section 3. Then, we present our lipreading dataset in section 4. In Section 5, we present the conducted experiments and obtained results. Section 6 summarizes our findings and outlines directions for future research.

Daou, S., Rekik, A., Ben-Hamadou, A. and Kallel, A.
Near-infrared Lipreading System for Driver-Car Interaction.
DOI: 10.5220/0011692300003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 631-638
ISBN: 978-989-758-634-7; ISSN: 2184-4321

2 RELATED WORK

Automatic visual speech recognition systems are classified into two types: word-level classification systems and character-level classification systems. In the context of human-machine interaction, we are interested only in word-level prediction, since the instructions provided by a car driver are limited to short and specific sentences and individual words dictionary. In this section, we present current advances in automatic lipreading systems based on deep learning for word-level prediction, starting with an overview of the available short-vocabulary lipreading datasets.

2.1 Short-Vocabulary Lipreading Datasets

The goal of automatic speech recognition systems is to understand natural speech, mainly structured in terms of sentences, which has made it necessary to acquire databases containing phonetically balanced words, phrases and sentences (Fernandez-Lopez and Sukno, 2018).

Among the formerly available datasets, we find VIDTIMIT (Sanderson, 2002), which was originally designed for people identification. It consists of 43 subjects uttering 10 sentences chosen among 346 different sentences. Similarly, AV-TIMIT (Hazen et al., 2004) was published in 2004 for audiovisual voice recognition. It contains 233 speakers and 510 different sentences. The audio-visual datasets that are accessible in English are summarized in Table 1. All of these datasets provide only RGB image sequences. There is currently no existing dataset that has been generated to address light restrictions.

2.2 Overview on Lipreading Methods

Due to the availability of extensive datasets and the advancement of deep learning techniques, there has been a considerable increase in the number of papers addressing the lipreading task during the last decade.

Gutierrez *et al.* (Gutierrez and Robert, 2017) presented a variety of models for predicting words using MIRACL-VC1 dataset (Rekik et al., 2014). They pre-processed the data by detecting and cropping the subject's face region in each video frame, and then concatenated the sequence of frames as input to their model. They investigated deep layered CNN baseline models additionally to LSTM network, inspired by Deep Mind's LipNet (Assael et al., 2016). The obtained results demonstrated the effects of dropout, hyperparameter setting, data augmentation, seen versus unseen validation partitions, batch normalization on adjusting these models.

Later, Stafylakis *et al.* (Stafylakis and Tzimiropoulos, 2017) developed a deep neural network for word-level visual speech recognition. It consists of a 3D convolutional neural network followed by a residual network that extracts more relevant visual representations as input at every time step to a two-layer Bidirectional Long Short-Term Memory. The word labels were repeated at every time step so that the overall loss is defined as the sum of losses across all time steps. Some variations of the network were investigated and trained in an end-to-end scheme on the LRW dataset (Chung and Zisserman, 2017). The best configuration improved the word prediction performance, achieving 83.0% of accuracy on the LRW dataset over previous works presented in (Chung and Zisserman, 2017; Chung et al., 2017).

Another prominent work is the multi-tower structure proposed by Chung and Zisserman in (Chung and Zisserman, 2018), where each tower takes a single frame or a T-channel image as input with each channel corresponding to a single frame in gray-scale. Then, the activation outputs from all the towers are concatenated to produce the final representation of the entire sequence. This multi-tower structure has been proved to be effective with appealing results on the current challenging dataset LRW.

For recognizing isolated words, (Ma et al., 2021) proposed a lipreading model that consists of 3D convolutional network similar to (Stafylakis and Tzimiropoulos, 2017) is followed by 18 layers of residual network and a temporal convolutional network (TCN). It achieved high performance on LRW and LRW-100 datasets, which are the largest publicly available datasets for isolated English word recognition task that outperforms all previous similar works. More recently the same authors (Martinez et al., 2020) introduced a novel depth temporal convolutional layer TCN head that reduces the computational cost. This proposed architecture outperforms the state-of-the-art performance with an accuracy of 88.6% and 46.6% on LRW and LRW-1000 datasets, respectively.

To achieve state-of-the-art performance, the vast majority of modern deep learning approaches require massive amounts of data, and their success in smaller datasets has been limited. As a result, some researchers claim that deep learning methods struggle with simple tasks and small-scale datasets (Petridis et al., 2020). We notice that most of the deep learning-based methods attempt to extract relevant visual features directly from input RGB frame sequences, rather than leveraging geometric features that can typically be extracted from mouth region and

Table 1: Short-vocabulary Lipreading Databases.

Name	Year	Cites	Language	Speakers	classes	Utterances
IBMViaVoice (Neti et al., 2000)	2000	312	English	290	10,500	24,325
VIDTIMIT (Sanderson, 2002)	2002	51	English	43	346	430
AV-TIMIT (Hazen et al., 2004)	2004	120	English	233	510	4,660
AVICAR (Lee et al., 2004)	2004	164	English	86	1317	59,000
OuluVS (Zhao et al., 2009)	2009	196	English	20	10	1,000
LILiR (Lan et al., 2010)	2010	60	English	12	200	2,400
UNMC-VIER (Wong et al., 2011)	2011	8	English	123	12	2,460
MOBIO (McCool et al., 2012)	2012	157	English	150	-	-
Austalk (Estival et al., 2014)	2014	8	English	1000	59	59,000
MIRACL-VC1 (Rekik et al., 2014)	2014	59	English	15	10	1,500
RM-3000 (Howell and Baker, 2015)	2015	4	English	1	1,000	3,000
OuluVS2 (Anina et al., 2015)	2015	32	English	53	530	530
TCD-TIMIT (Harte and Gillen, 2015)	2015	46	English	62	5,954	6,913
IBM AV-ASR (Mroueh et al., 2015)	2015	72	English	262	10,400	-
AV Digits (Petridis et al., 2018)	2018	2	English	39	10	5,850

deformations. In this paper, we propose to design a two-stream deep learning architecture that combines both geometric and global visual features extracted from the mouth region to improve the performance of lipreading based only on visual cues.

3 METHODS AND MATERIALS

3.1 Proposed Approach

As shown in figure 1, the proposed system has three main stages. The first stage is dedicated to preprocessing the input video and extracting relevant facial information like mouth region and landmarks. In the second stage, we used a two-stream feature encoder module consisting in a global feature network that aims to model the global motion information of the mouth area to get comprehensive information related to the visual speech, and a 2D lips landmarks module to encode lip contour information and local motion information around the lip. Finally, the computed features from the output of the encoder are concatenated and sent into a temporal model to capture the temporal dependencies followed by a softmax layer to compute the class probabilities for different commands in our system.

3.1.1 Video Preprocessing

The goal of this step is to reduce the effect of the face pose variation in different video frames. First, 68 facial landmarks are detected in all frames in the input video using a facial landmark detection algorithm (Sagonas et al., 2013). Then, each face frame is aligned to a reference mean face shape. Finally, the mouth area is cropped from the aligned face frames so that the mouth region is always roughly centered on the image crop. In this stage, facial landmarks are only used to determine the mouth location, to align faces and to crop the mouth region.

3.1.2 Visual Front-End Network

Global-Feature Network. The aim of the global feature network is to encode global characteristics of lip movements on the cropped mouth region. Inspired from (Ma et al., 2022), this network consists on a 3D convolutional layer, which takes as an input $W \times H \times T$ consecutive frames followed by a 2D ResNet-18 (Stafylakis and Tzimiropoulos, 2017).

Landmark-Feature Network. The detected facial landmark in the preprocessing step aims to determine the locations of significant facial points describing the unique location of a facial component (eye corner, mouth corner) or an interpolated point connecting those points (Wu and Ji, 2019). In our system, only 33 facial landmark points from the mouth area are selected as lipreading-related landmarks. These facial landmarks are then encoded using one layer graph convolutional network (GCN) (Kipf and Welling, 2016). Each frame is represented as a graph node and every node is related to the two nearest nodes. In other words, every frame is related to the previous frame and to the next frame of the video sequence. The input feature dimension at the node level is 33×2, corresponding to the lip landmark coordinates for each

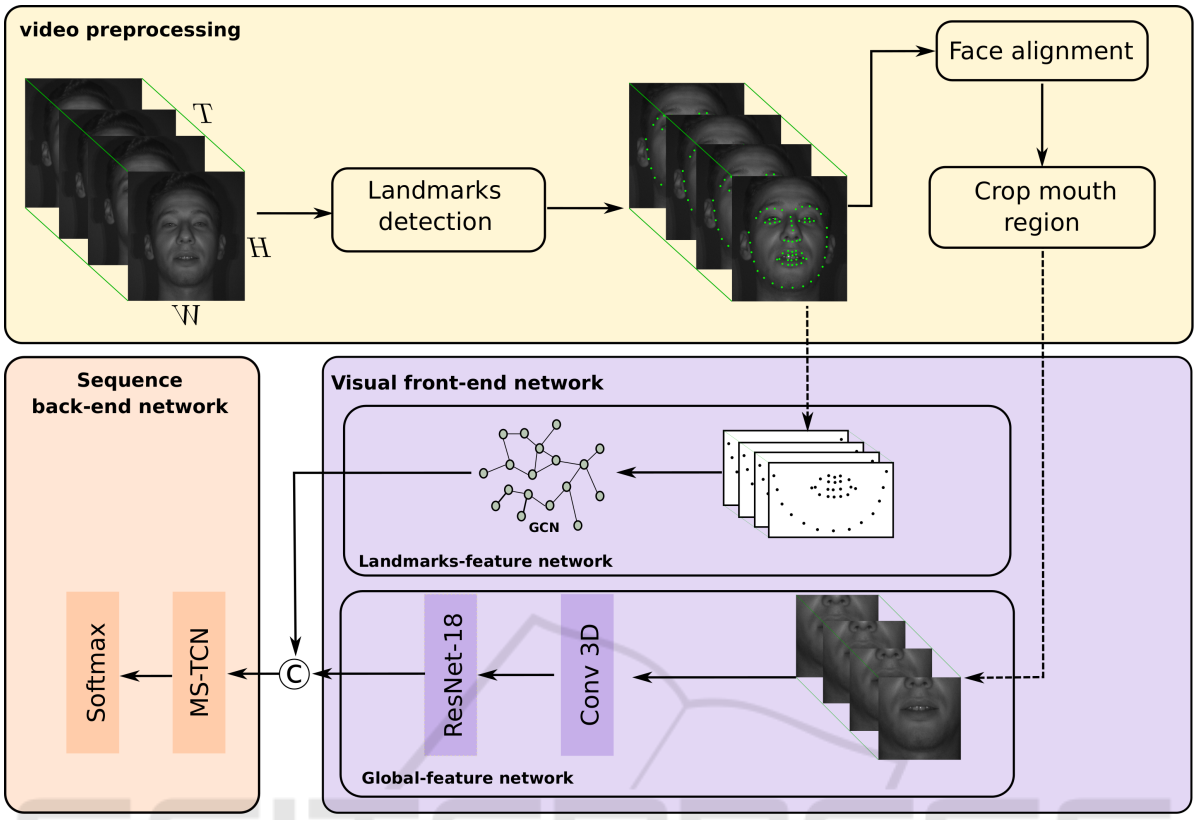


Figure 1: Three stages framework of the proposed system. Video preprocessing: extract facial landmarks and crop the mouth region from the input infrared video. Visual front-end network: encode global visual and lips contour variation on the cropped mouth area. Sequence back-end network: based on a multi-scale temporal convolutional network (MS-TCN) to encode temporal variation along the extracted features and classify the input video.

frame, resulting in a 512 dimensional output feature vector.

3.1.3 Sequence Back-End Network

The goal of this network is to map the landmark-features and the global-features extracted from the visual front-end network. Our back-end network is based on the temporal convolution network (TCN) since it improves considerably the performances on word-level lipreading tasks (Ma et al., 2020). The proposed model consists of Multi-Scale dilated TCN layers, a fully connected layer, and a final softmax layer. In this variant, each TCN layer consists of several branches with different kernel sizes. Figure 2 presents a detailed representation of this network architecture.

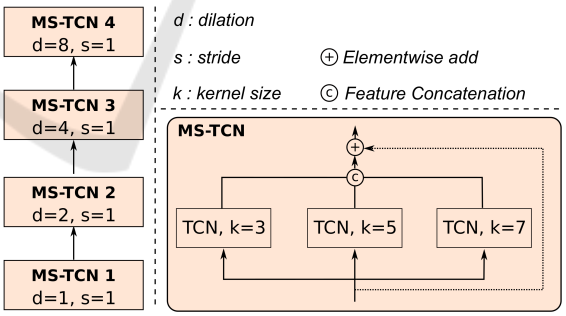


Figure 2: Sequence back-end network: it contains four layers of Multi-scale dilated TCN, where the dilation size of each layer is 1, 2, 4, and 8, respectively. Each layer consists of three TCN blocks with kernel sizes 3, 5, and 7, respectively.

3.2 Lipreading Infrared Dataset

This section describes our multi-step pipeline for collecting and processing our dataset of near-infrared lipreading in driving mode. We call this dataset infrared-LR. 29 speakers were involved to obtain a global number of 1044 utterances. Each speaker repeated 12 representative car commands three times, which are previously selected in collaboration with a car maker partner (see the command list in Table 2). Near-infrared cameras are usually used in the car cockpit since they can efficiently capture the car interior in both day and night conditions. In night condi-

tions, the device registers with remote infrared LEDs or variable range that activate when outside lighting is inadequate. The voice is also recorded alongside the videos for further investigations.

Table 2: Database dictionary.

Time to arrival
Weather forecast
Cooler
Warmer
Take me home
Take me to work
Take a selfie
I feel fine
I need a break
Mute
Accept call
Reject call

The voice recordings are also used to establish a temporal alignment of the spoken audio with text transcription, and then to construct a spatial-temporal alignment for the frames matching to the word sequence. Figure 3 summarizes the pipeline, and the various processes are detailed in the following paragraphs.

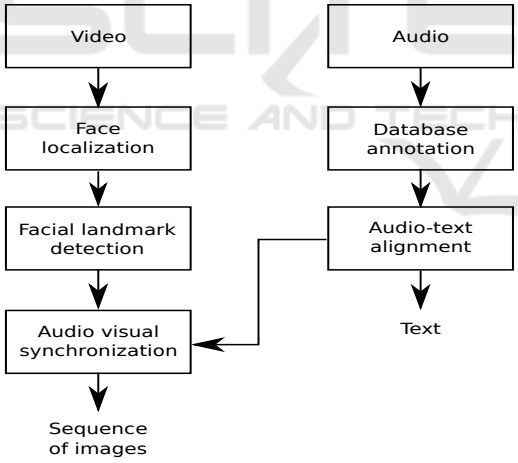


Figure 3: Database generation pipeline.

Subjects. The automotive vehicle instructions were uttered by 29 speakers, 19 male and 10 female, ranging in age from 18 to 40 years. Figure 4 presents a sample of speakers from the infrared-LR dataset.

Database Annotation. We tested standard voice recognition APIs to annotate the sequences, however, it was not that efficient. Although, the recognition rates were acceptable, the generated starting and ending timestamps were mostly imprecise and system-

Figure 4: Samples from infrared-LR dataset.

atically required manual correction. This leads to switching immediately to manual annotation. The text-to-audio alignment is done semi-automatically using a mini-script that includes PyQt5 features (as shown in Figure 5).

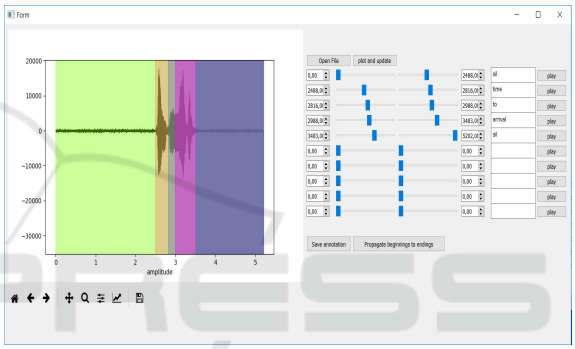


Figure 5: The developed GUI to check the automated annotation. If the automated annotation fails for some reason, the GUI allows for modifying both text labels and speech intervals.

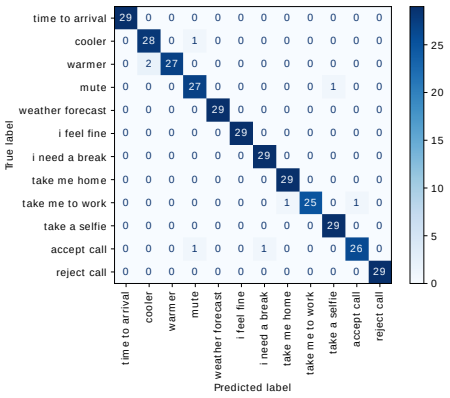
Face Localization. The HOG-Based DLIB face detector is used to identify facial appearances in each video frame. Using a KLT detector, all face detections are then sorted into face tracks.

Facial Landmarks Detection. To identify the mouth location, facial landmarks are required. They are derived from the iBug face landmark predictor (Sagonas et al., 2013), which has 68 landmarks. Using these landmarks, we perform an affine transformation to obtain a mouth-centered crop of 88×88 pixels per frame.

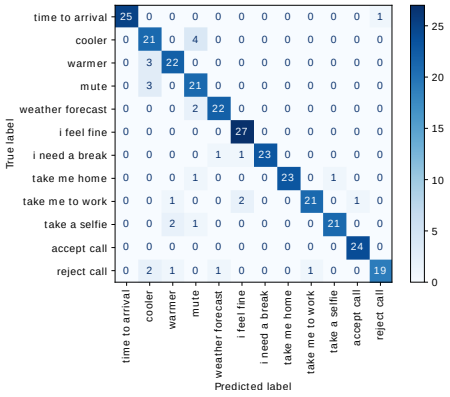
4 EXPERIMENTS

4.1 Experimental Settings

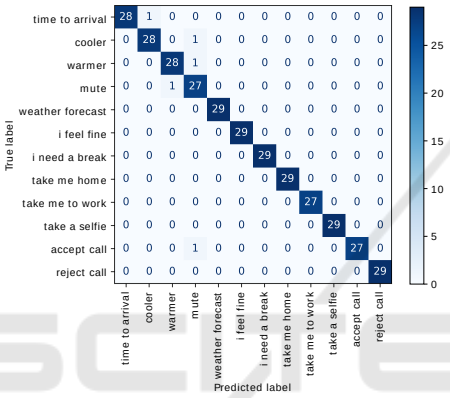
Our implementation is based on the pyTorch library (Paszke et al., 2019). The proposed architecture is



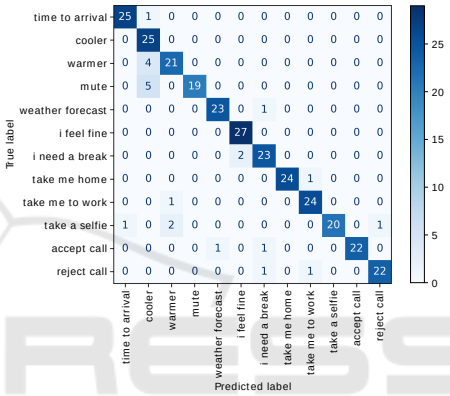
(a) SD without landmark features.



(b) SI without landmark features.



(c) SD with landmark features.



(d) SI with landmark features.

Figure 6: Obtained confusion matrices for different configurations. a, b: speaker-dependent and speaker-independent configuration, respectively, using only the global-feature network. c, d: speaker-dependent and independent configuration, respectively, using both landmark-feature and global-feature networks.

trained for 200 epochs on an NVIDIA Titan V GPU with 12GB memory, with a mini-batch size of 8. The AdamW optimizer (Loshchilov and Hutter, 2017) is used, with an initial learning rate of 3e-4. The learning rate is decayed without a warm-up phase using a cosine annealing strategy. For all experiments, we also employ variable-length augmentation (Martinez et al., 2020).

4.2 Experimental Results

As a first attempt toward developing a lip-reading system that it suitable to driver-car interaction, we start by running a series of experiments to assess the proposed system's performance on our infrared-LR dataset.

The proposed system is evaluated on two configurations: subject-dependent (SD) and subject-independent (SI). For SD Configuration, all speakers' videos are used for both the training and the validation stages. However, for the SI configuration, the separation of training and validation data is done at the speaker level where speakers present in the training data are not present in the validation subset.

To demonstrate the importance of combining landmark features and global visual features, we evaluate our system using global features only and using global features combined with landmarks features for both SD and SI configurations. Table 3 presents the obtained results for the different configurations. The values presented in this table correspond to the rate of the lip-reading system's accuracy, which indicates the proportion of instructions accurately predicted relative to all commands delivered in the test set of data.

Table 3: Obtained lipreading performance for the different setting combinations.

Configuration	Landmarks (-)	Landmarks (+)
SD	97.6%	**98.5%**
SI	90.2%	**92.2%**

Experiments for the SD Configuration. The training/testing split is as follows. As each command is uttered 3 times for each user, we choose to randomly select 2 sequences for training and use the remaining one for validation. As a result, 29×12 sequences are used for the validation. The overall obtained recognition rate is equal to 97.6% without considering the landmark features as additional data, while the full model achieved a high performance of 98.5%.

Experiment for the SI Configuration. A performance drop is usually expected when comparing the SI configuration to the SD configuration. Indeed, the overall performance obtained for the SI configuration is 92.28%, however, it is equal to 90.2% without using landmark features.

We present also the obtained confusion matrices for the different configuration combinations (see figure 6). Ideally, the diagonal entries of a confusion matrix are equal to one and zeros everywhere else. Globally, we can observe this trend in the obtained confusion matrix. We can also observe a relatively important confusion between the short commands "*mute*", "*cooler*", and "*warmer*". These commands have roughly speaking the same length and induce the same visible lips movements, which explains this confusion.

5 CONCLUSION

In this paper, we proposed a lipreading system for driver-vehicle interaction based on near-infrared cameras. The approach is based on a deep learning architecture with two-stream network architectures that combines both geometric and global visual features extracted from the mouth region to improve the performance of lipreading based only on visual cues. Also, we constructed the first near-infrared lipreading dataset for driver-car interaction, named infrared-LR. The experimental results show, for both speaker-dependent and speaker-independent configurations, that our hybrid design produced a high performance with a considerable improvement over applying only the global-feature network.

REFERENCES

Afouras, T., Chung, J. S., and Zisserman, A. (2018). Deep lip reading: a comparison of models and an online application. *arXiv preprint arXiv:1806.06053*.

Anina, I., Zhou, Z., Zhao, G., and Pietikäinen, M. (2015). Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–5. IEEE.

Assael, Y. M., Shillingford, B., Whiteson, S., and De Freitas, N. (2016). Lipnet: Sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2(4).

Ben-Hamadou, A. (2020). Control method, control device, system and motor vehicle comprising such a control device. US Patent 10,627,898.

Chung, J. S., Senior, A. W., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In *CVPR*, pages 3444–3453.

Chung, J. S. and Zisserman, A. (2017). Lip Reading in the Wild. In Lai, S.-H., Lepetit, V., Nishino, K., and Sato, Y., editors, *Computer Vision – ACCV 2016*, volume 10112, pages 87–103. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Chung, J. S. and Zisserman, A. (2018). Learning to lip read words by watching videos. *Computer Vision and Image Understanding*.

Estival, D., Cassidy, S., Cox, F., Burnham, D., et al. (2014). Austalk: an audio-visual corpus of australian english. In *Proceedings of the International Conference on Language Resources and Evaluation*. Reykjavik, Iceland: European Language Resources Association.

Fernandez-Lopez, A. and Sukno, F. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*.

Gutierrez, A. and Robert, Z. (2017). Lip reading word classification.

Harte, N. and Gillen, E. (2015). Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615.

Hazen, T. J., Saenko, K., La, C.-H., and Glass, J. R. (2004). A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 235–242. ACM.

Howell, A. and Baker, L. (2015). *Confusion Modelling for Lip-Reading*. PhD thesis, School of Computing Sciences. University of East Anglia.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lan, Y., Theobald, B.-J., Harvey, R., Ong, E.-J., and Bowden, R. (2010). Improving visual features for lipreading. In *Auditory-Visual Speech Processing 2010*.

Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., and Huang, T. (2004). Avicar: Audio-visual speech corpus in a car environment. In *Eighth International Conference on Spoken Language Processing*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, P., Martínez, B., Petridis, S., and Pantic, M. (2020). Towards practical lipreading with distilled and efficient models. *CoRR*, abs/2007.06504.

Ma, P., Martinez, B., Petridis, S., and Pantic, M. (2021). Towards practical lipreading with distilled and efficient models. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7608–7612. IEEE.

Ma, P., Wang, Y., Petridis, S., Shen, J., and Pantic, M. (2022). Training strategies for improved lip-reading. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8472–8476. IEEE.

Martinez, B., Ma, P., Petridis, S., and Pantic, M. (2020). Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE.

McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matejka, P., Cernockỳ, J., Poh, N., Kittler, J., Larcher, A., Levy, C., et al. (2012). Bi-modal person recognition on a mobile phone: using mobile phone data. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 635–640. IEEE.

Mroueh, Y., Marcheret, E., and Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2130–2134. IEEE.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., and Mashari, A. (2000). Audio visual speech recognition. Technical report, IDIAP.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Petridis, S., Shen, J., Cetin, D., and Pantic, M. (2018). Visual-only recognition of normal, whispered and silent speech. *arXiv preprint arXiv:1802.06399*.

Petridis, S., Wang, Y., Ma, P., Li, Z., and Pantic, M. (2020). End-to-end visual speech recognition for small-scale datasets. *Pattern Recognition Letters*, 131:421–427.

Rekik, A., Ben-Hamadou, A., and Mahdi, W. (2014). A new visual speech recognition approach for rgb-d cameras. In *International conference image analysis and recognition*, pages 21–28. Springer.

Rekik, A., Ben-Hamadou, A., and Mahdi, W. (2015a). Human machine interaction via visual speech spotting. In *Advanced Concepts for Intelligent Vision Systems*, pages 566–574. Springer.

Rekik, A., Ben-Hamadou, A., and Mahdi, W. (2015b). Unified system for visual speech recognition and speaker identification. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 381–390. Springer.

Rekik, A., Ben-Hamadou, A., and Mahdi, W. (2016). An adaptive approach for lip-reading using image and depth data. *Multimedia Tools and Applications*, 75(14):8609–8636.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403. IEEE.

Sanderson, C. (2002). The vidtimit database. Technical report, IDIAP.

Sheng, C., Kuang, G., Bai, L., Hou, C., Guo, Y., Xu, X., Pietikäinen, M., and Liu, L. (2022). Deep learning for visual speech analysis: A survey. *arXiv preprint arXiv:2205.10839*.

Stafylakis, T. and Tzimiropoulos, G. (2017). Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*.

Sun, K., Yu, C., Shi, W., Liu, L., and Shi, Y. (2018). Lip-interact: Improving mobile device interaction with silent speech commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 581–593.

Wong, Y. W., Ch'ng, S. I., Seng, K. P., Ang, L.-M., Chin, S. W., Chew, W. J., and Lim, K. H. (2011). A new multi-purpose audio-visual unmc-vier database with multiple variabilities. *Pattern Recognition Letters*, 32(13):1503–1510.

Wu, Y. and Ji, Q. (2019). Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142.

Zhao, G., Barnard, M., and Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265.