Lip Reading in Greek words at unconstrained driving scenario

Dimitris Kastaniotis
Computer Vision Systems,
IRIDA Labs S.A.
Patras, Greece
dkastaniotis@iridalabs.gr

Dimitrios Tsourounis
Department of Physics
University of Patras
Patras, Greece
dtsourounis@upatras.gr

Aristotelis Koureleas
Department of Physics
University of Patras
Patras, Greece
aristoteliskoureleas@gmail.com

Bojidar Peev
Department of Physics
University of Patras
Patras, Greece
up1048684@upnet.gr

Christos Theoharatos
Computer Vision Systems,
IRIDA Labs S.A.
Patras, Greece
htheohar@iridalabs.gr

Spiros Fotopoulos
Department of Physics
University of Patras
Patras, Greece
fotopoul@upatras.gr

***Abstract*—This work focuses on the problem of Lip Reading with Greek words in an unconstrained driving scenario. The goal of Lip Reading (LR) is to understand the spoken work using only visual information, a process also known as Visual Speech Recognition (VSR). This method has several advantages over Speech Recognition, as it can work from a distance and is not affected by other sounds like noise in the environment. In this manner, LR can be considered as an alternative method for speech decoding which can be combined with state-of-the-art speech recognition technologies. The contribution of this work is two-fold. Firstly, a novel dataset with image sequences from Greek words is presented. In total, 10 persons spoke 50 words while they were either driving or simply sitting in the passenger's seat of a car. The image sequences were recorded with a mobile phone mounted on the windshield of the car. Secondly, the recognition pipeline consists of a Convolutional Neural Network followed by a Long-Short Term Memory Network with a plain attention mechanism. This architecture maps the image sequences to words following an end-to-end learning scheme. Experimental results with various protocols indicate that speaker independent Lip Reading is an extremely challenging problem.**

Keywords— Lip-Reading, Visual Speech Recognition, Deep Learning, Speech decoding.

I. INTRODUCTION

Lip Reading (LR) or Visual Speech Recognition (VSR) is the task were the spoken word is perceived only by visually observing the motion of the mouth [1]–[3]. In this context, lip reading is a special problem of human action understanding, were the information is observed in human body movements [4]. Although people are very good in understanding and extracting high level description from a biological human motion like action, gestures and gait, they do not perform well on the task of recognizing the spoken word by visually observing the motion of the mouth. Indeed, the estimated human-level performance on Lip Reading is poor. In particular, it has been observed that humans achieve 17% and 21% for 30 monosyllabic and compound words respectively [5]. The poor performance is related with the homophemes (different characters that produce similar motion) as well as the fact that people heavily rely on the sound in order to perceive a spoken word [6].

Deep Convolutional Neural Networks (Deep CNNs) have achieved state-of-the-art performance in most computer vision problems, since they can extract robust visual feature representations based on an end-to-end learning process from a large amount of data [7]. Thus, the use of deep architectures, can boost the recognition systems' efficiency in the LR problem and even surpass the human-level performance [8]. For example, the authors in [5] achieved a performance of 91.4% and 93.2% on a dataset with phrases and 61.1% on a total number of 500 words.

Whilst these methods achieve very high accuracies, it is worth to study how these techniques generalize well to large number of people, large number of words and different languages. More specifically, it is mandatory to evaluate the performance on a speaker-independent way [9] because in many tasks, the biological human motion encompasses information about the user's identity. In that case, the LR systems could benefit from the person's regularities in order to fictitiously improve their efficiency. Also, regarding the spoken languages, most studies focus on one language and in this manner, it is also important to investigate what would be the performance in languages with different phonemes.
This work tries to study the LR problem towards the following directions.

Firstly, by presenting a new dataset with Greek words recorded with a mobile phone mounted on the frontal car windshield in an unconstrained driving scenario. This dataset consists of 10 persons who speak 50 words repeatedly for 5 times. To the authors' knowledge, this is the first time a database for LR about Greek words is being publicly available. This dataset can help in exploiting the performance of LR in Greek language and also, analyzing the performance of LR between different languages. The dataset along with accompanying code and information will be made free-available in the following project page: https://github.com/dimkastan/LipReadingGreekWords.

Secondly, the experimental results focus on the ability of the system to generalize at new persons. This speaker-independent evaluation is of major importance in order to interpret properly Lip Reading capabilities as well as limitations. For the classification of input image sequences, a combination of a Convolutional Neural Network (CNN) followed by a Long-Short Term Memory Network (LSTM) and a plain attention mechanism, which averages all LSTM responses, is utilized. The aim of this architecture is to learn the mapping of an image sequence of a word to the corresponding word class.

II. Related Work

Word level automated Lip Reading is the process where a machine is learning the mapping between the input image sequence and a spoken word. In order to achieve this, traditional methods utilized low-level descriptors such as Local Binary Patterns, trajectories of the distances between particular points of the mouth and spatiotemporal features based on Active Appearance models [6]. In the modeling of sequential information, Hidden Markov Models (HMM) were the dominant approach [10]. In other approaches, gradient based features using Histograms of Oriented Gradient (HOG) as well as Motion Boundary Histograms (MBH) along with Support Vector Machines (SVMs) were used [11].

These methods were characterized by two main limitations. Firstly, the hand crafted features were characterized by poor representational performance [5], [8], [12]. In this manner, they were highly affected by a variety of factors like lighting. Secondly, the sequence modeling with HMM resulted into poor sequence representation performance. With the emergence of data driven learning and in particular, with end-to-end learning architectures, which are able to learn high level feature representations, the problem of lip reading was reconsidered. Also, in this direction, the use of Recurrent Neural Networks (RNNs), which have the ability to take advantage of sequence data, boosted the performance of existing systems significantly [9].

As the problem of lip reading is trying to learn the mapping of a sequence with arbitrary frames to a finite number of spoken words, it is developed utilized recurrent learning modules [2] or the input sequence is treated as fixed length by padding the frames [8]. More specifically, in [12] a combination of a CNN with an LSTM Neural Network was proposed for the task of lip reading. In [8], frames were processed by a CNN in two different fusion schemes in order to produce a fixed length representation.

Of major importance is also the availability of publicly available datasets. These datasets allow and standardize the comparison between different methods for LR. In this context, several datasets have been presented covering different aspects of the problem as for example are the large number of words and the large number of persons. More specifically, the MIRACL-VC1 dataset [11] provides 10 words from a total number of 15 persons in a constrained environment were each word is repeated 10 times. The authors in [11] used a speaker dependent and a speaker independent evaluation protocol. Also, it is worth noticing that the performance is significantly reduced during the above speaker independent evaluation [11]. A second dataset which provides a large number of words spoken by 1000 subjects is presented in [8]. These data are collected automatically from the web and the words are cropped using information from the subtitles. Whilst authors mention that they provide a dataset for speaker independent evaluation it is not clear how the distribution of the words per subject (person) could potentially affect the performance. Also, as all words are taken from TV news, it is obvious that the scenario is far from unconstrained as the lightning conditions, the camera view and the reflections are fully controlled. By taking into account all these issues, here is presented a dataset for evaluating LR problems with Greek words.

III. Dataset

In order to evaluate the performance of automated LR in Greek words in unconstrained scenarios as well as to promote a speaker independent evaluation protocol, a novel lip reading dataset was generated. For this purpose, a total number of ten native Greek speaking persons (4 females and 6 males) were used. Each person was sitting in the front seats of a car as a driver or a passenger. If the person was not driving, he was advised to watch the environment imitating as close as possible the driver's behavior. In cases were the car was parked, the subject was advised to move naturally and look outside in order to observe the environment. The camera was mounted internally of the frontal windshield. Between words there was a pause and the videos were semi-automated cropped by adding a sound of a specific frequency between each spoken word. The recordings performed in unconstrained scenario (lighting, viewpoint) in different cars and from different mobile phones. In total 50 Greek words have been recorded in this version of the dataset. These words form a common lexicon related to everyday driving and transportation activities plus some extra possible navigation commands, like possible locations (e.g. University, Workplace, Gas Station, etc.) and functionalities related to the car operation (e.g. alarm, wipers, etc.). The complete list of the words as well as the code for reproducibility will be available with the complete dataset after publication in the following Git-Hub project page: https://github.com/dimkastan/LipReadingGreekWords. Table I presents a summary of the dataset.

TABLE I. Summary of dataset characteristics

Video Resolution	1280x720
Number of Persons	10
Number of words (categories)	50
Samples per word and per person	5
Sequence Length (mean±std)	44 ± 15
Audio Recordings	Yes
Frames per Second	25

The words were cropped semi-automatically. More specifically, between every spoken word, a sound produced by a clap of the hands was used in order to separate the videos into small clips containing the words for the next cropping process. Then, the videos were visually inspected and classified to one of the categories by a person (data annotation). During the annotation, some bad samples were rejected. Finally, for each word, a sequence of frames and a small audio clip are produced. For the LP task only the image sequences are used. As a post processing, the frames of every sequence are being fed into a processing pipeline in order to detect and crop the area of the mouth. In this pipeline, initially the face is being detected using a face detector based on a cascade classifier and then the mouth region was detected by finding the fiducial points around the mouth using a very fast algorithm [13].

The dataset captured with totally different lighting conditions and view angles (between users). In Fig. 1 a sequence captured from a subject while driving is presented. By observing the background, the view angle and the lighting

variations on the cropped mouth patches, it is clear that the particular task is very challenging.

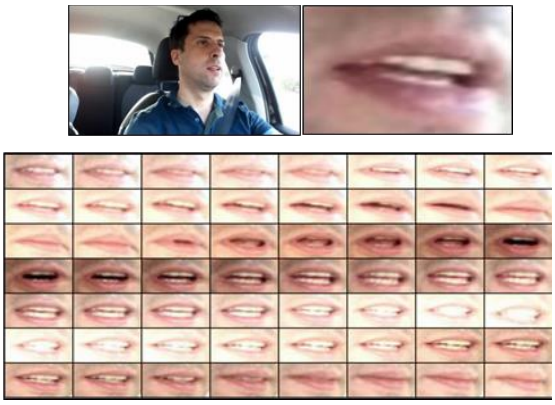

Fig. 1. Top Left: A random frame from a word sequence. Right: The mouth patch of the face. Bottom: A sequence of a spoken word (from top to bottom and from left to right). The lighting variations on the cropped mouth patches are caused by sunlight reflections which vary significantly even within the time window of a single word.

IV. Proposed Framework

A. Overview

The proposed method is based on a combination of a Convolutional Neural Network followed by a bidirectional Long-Short Term Memory Network (LSTM) and two fully connected layers followed by a plain attention mechanism. More specifically, the CNN takes as input the area of the mouth and produces a vector representation for each single frame of the word sequence. These vectors are then fed into a bidirectional LSTM, which encodes the sequence into a vector following a many-to-one mapping. The architecture of the proposed method is presented in the following Fig. 2.

B. Mouth detection

Given a sequence of frames captured during the pronounce of a word, initially, and for all frames of the sequence, the mouth patches are been detected and cropped as shown in Fig. 2-1. The mouth is detected as follows. First, the fiducial points of the face are been detected. From a total number of 65 fiducial points, 20 points correspond to the mouth. Then, a bounding rectangle increased by 10% in width and height is calculated. The patch of the mouth is then cropped in arbitrary size.

C. Feature extraction from mouth patches

The patches of the mouth that extracted from the sequence frames are fed into a Convolutional Neural Network (CNN) which produces a vector representation of the input mouth patches as shown in Fig. 2-2. More specifically, as the size of the input patch varies significantly, here it was selected to resize all input patches to a fixed size of 64 by 128 (width and height) respectively. The CNN used here is a ResNet-18 [7], which produces an 512-dimensional feature map. The feature map is then processed by global average pooling resulting into a 512- dimensional feature vector. The weights of the network are initialized from a pretrained model trained on Image classification [7] or randomly, depending on the experiment.

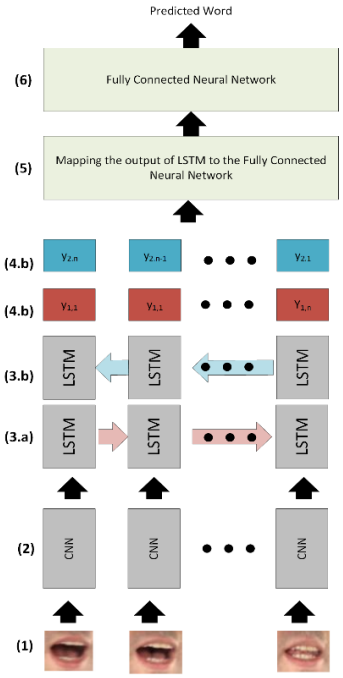

Fig. 2. Overview of the proposed method. Each image of the sequence is mapped to a feature vector using a CNN. These vectors are fed into a Bidirectional LSTM and provide a new mapping. The attention mechanism encodes the distribution of the LSTM responses and feeds a Fully Connected Network.

D. Sequence encoding

Given a sequence of feature vector, the goal is to learn the mapping of the sequence to a fix number of words (categories). In order to achieve this, a recurrent learning architecture, which can learn long range time dependencies, is utilized. This network will provide an encoding of the sequence into a fixed length vector. In particular, the feature vectors are fed into a Long-Short Term Memory (LSTM) Neural Network, as shown in Fig. 2-3. The LSTM consumes the sequences of the vectors and produces another sequence of the vectors represented as $y_t \in \mathbb{R}^{256}$ as shown in Fig. 2-4. For the task of lip reading, where the whole sequence is available for evaluation and the task is to predict the spoken word, the sequence is fed into the LSTM network in both directions - namely forward and reverse. At each time step, the LSTM is producing a feature response. A common approach is to final response of the LSTM following a many to one approach. An alternative approach, which also allows the CNN to be trained jointly with the LSTM is to apply a cost function at each time step [12]. Here, all responses were fused by averaging them (Fig. 2-5). Thus, in this work, the responses of the LSTM for each time *t* if the sequence of length T, were averaged using formula (1).

$$y_o = \frac{\sum_{t=1}^{T} y_t}{T} \quad (1)$$

The previous formula belongs to a special type of attention mechanism [14] where each time step contributes equally to the weighted sum as shown in formula (2). The weight w_t is a scalar equal to 1/T, where T equals the time-steps of the sequence. The symbol · denotes a multiplication between a scalar and a vector.

$$y_o = \sum_{t=1}^{T} (w_t \cdot y_t) \quad (2)$$

E. Classification and learning criterion

The encoded sequence is then fed into a Fully Connected Neural Network (FCNN) (Fig. 2-6). This FCNN, is mapping the encoded feature vector into one of the categories corresponding to the number of the words. The final mapping is performed using a Soft-Max response function, which provides probabilities for each response. As criterion, Cross Entropy Loss is used, which can be defined as:

$$-\sum_{i=1}^{C} y_{o,c} \log(p_{o,c}) \quad (3)$$

, where $y_{o,c}$ is a one-hot encoding vector whose dimensions correspond to the C categories, and $p_{o,c}$ is a vector with probabilities computed by the Soft-Max layer.

V. Experimental Results

The goal of the following experimental results is two-fold. Firstly, to assess the ability of automated lip reading in Greek words and secondly, to evaluate the generalization ability of the models. For this purpose, two evaluation protocols are proposed. The Leave-N samples out and the Leave-N persons out protocols. These protocols study the generalization ability of the model and the ability of the model to generalize to new users. In both cases, the random classification rate for 50 words is 2% (i.e. 1/50).

All experiments performed on Ubuntu 16.04 using the Python Programming Language. For the Deep Learning models PyTorch framework was used [15]. For face detection and fiducial point extraction (facial landmark detection), the Python interface of DLib library [16] was used. The models were trained for 100 epochs and the test is being reported for the model with the best validation accuracy. All networks were trained with the SGD optimizer with learning rate 0.01 with an exponential learning rate strategy, applied every 30 epochs. However, no more than 50 epochs were required in most cases.

A. Leave-N samples out protocol

In this protocol, a split of 60-20-20 % was followed. From a total number of 5 samples per word and per user, N=1 sample was kept out as test and 1 sample was used for validation. The rest 3 samples were used for training the network. The results reported in the following tables correspond to the test set. The demonstrated test results are reported for the model that achieved the best validation set accuracy. Results indicate that bidirectional LSTM with a pretrained backend CNN on Image Classification task achieve the best performance. Regular LSTMs with pretrained CNN have limited generalization performance. However, when the convolutional networks are trained from scratch (i.e. the weights of the ResNet-18 CNN are randomly initialized), the bidirectional LSTM seems to suffer from significant overfit (Table II). On the contrary, the use of regular LSTM (single directional) has higher performance but still, much lower than the models based on pretrained CNN weights. Thus, pretrained CNNs seem to improve the generalization performance of the proposed method.

TABLE II. Classification Results for Hold-out evaluation.

LSTM form	Pretrained ResNet-18	ResNet-18 from scratch
Bidirectional	59.96 %	23.24 %
Single Direction	50.58 %	39.06 %

B. Leave-N-Persons out Protocol

While the performance of the Leave-N samples out is very satisfying, and in accordance with the results reported by other state-of-the-art research works, it is important to study the ability of the system to generalize well to new users. This procedure provides a speaker independent evaluation, which is critical for the evaluation of the method since in this kind of problems most approaches tend to learn user-dependent features [17]. This speaker independent scenario is of major importance and it is expected to unveil more insights that affect the performance of the lip reading system. In this manner, here, the following procedure is followed. From the total number of 10 persons, 6 are used for training, 2 for validation and 2 for test.

Results are being reported in the following Table III as mean± standard deviation. The experiment was performed 5 times with randomly selected test persons (N=2). Results indicate that Bidirectional LSTMs with pretrained weights generalize better as compared to the other combinations, something which is also expected by the findings of Table II.

TABLE III. Leave Two persons out protocol. In total 6 persons are used for training, 2 for validation and 2 for test.

	Classification Accuracy (%)	
LSTM form	***Pretrained ResNet-18***	***ResNet-18 from scratch***
Bidirectional	14.86 ± 2.83	11.38 ± 0.89
Single Direction	11.65 ± 1.42	10.09 ± 0.97

C. Classification Accuracy with respect to the number of words

In this section, the classification performance of the model is evaluated with respect to the number of the words. This is achieved by varying the total number of categories of the classification problem. In order to provide a thorough evaluation, the Leave-Two persons out as well as the model specifications of previous Section *B* is followed here.

For this purpose, several experiments were performed with the number of words varying from 2 to 50 (in particular 2, 5, 10, 20, 30 and 40 words are used respectively). Fig. 3 presents the box plots of 20 random repetitions (random selection of the words' categories) in this speaker-independent scenario. Thus, here, the words from two users in the classification performance decreases as the number of words increase. This is expected as the networks try to decode words with different mouth movements, various lengths and poses.

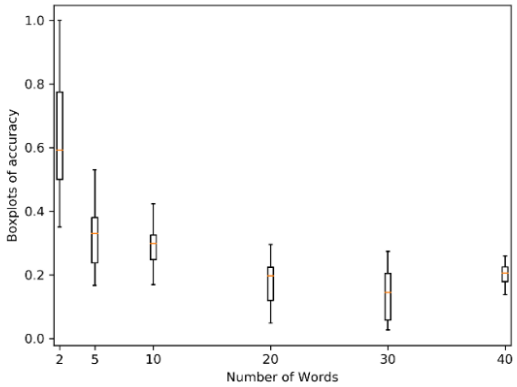

Fig. 3. Classification Accuracy (y-axis) with respect to the number of words (x-axis). The evaluation performed with the Leave-N-Persons-Out protocol. Here, two persons were used for test.

D. Model Attention

Visually inspection of speech image sequences is interesting, as it could provide some information about the spatial contribution of the pixels of the mouth region at each time step. Besides, this information can help in transferring the knowledge to other networks, even in generative ones [18]. Thus, the visualization of the CNN's activations, when one sequence is input, provides some information for the model's attention. In order to achieve this, the responses (only positive numbers) of all channels (here 512) in the last convolutional layer $F \in \mathbb{R}^{512x8x16}$, where $F_i \in \mathbb{R}^{8x16}, \forall \mathrm{i} = 1, \ldots, \mathrm{N}$, and N=512 equals the number of channels of the final convolutional layer feature map, are summed as shown in equation (4).

$$A = \sum_{i=1}^{N} F_i \tag{4}$$

The produced heatmap $A \in \mathbb{R}^{8x16}$ is then upscaled with bilinear interpolation in order to match the dimension of the input image. The produced heat-map is blended with the input image using alpha-blending. A better understanding of the CNN's behavior can be obtained, if the responses of a CNN trained on the task of lip reading (Fig. 4) are compared with those of a CNN trained on a different task (e.g. image classification). In Fig. 4, it is clear that the responses are very localized to the mouth, whilst in Fig. 5 the network's attention is more diffused, covering the whole mouth region as well as areas outside the mouth.

Fig. 4. Visualizing the Attention of a Convolutional Neural Network trained on the Lip-Reading dataset on a word sequence.

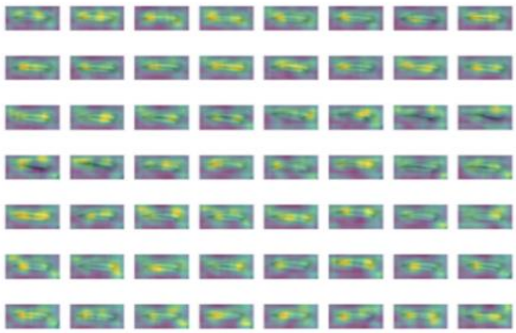

Fig. 5. Visualizing the Attention of a CNN trained on the task of Image classification.

Visualizing the attention of a CNN trained on LR could enlighten the ability of the CNN to transfer its knowledge into other languages. In particular, it is expected that CNNs try to learn a mapping that is related to specific phonemes. As phonemes change across languages, the transferability of knowledge could be very challenging and further research is required.

Also, by visualizing the attention, it is probably feasible to determine if the network is able to provide a robust feature representation, or it fails. In order to achieve this, a speaker independent scenario is evaluated. More specifically, a misclassified test sample is been considered (from a user never seen before in the training or validation set). By observing the attention map of the network's response (Fig. 6), it is obvious that the CNN has totally failed into detecting the areas of interest (meaning the areas that are related with specific phonemes).

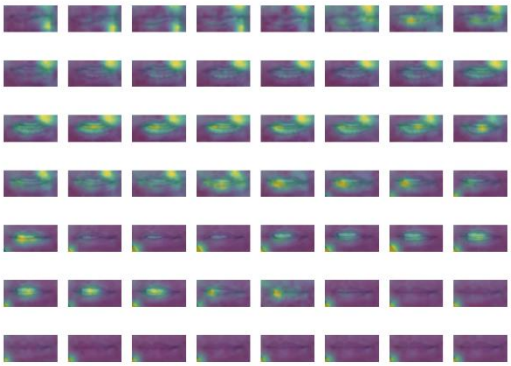

Fig. 6. The attention map for a misclassified sample from the test set under the speaker independent evaluation protocol. The network seems to fail to finely localize the mouth regions as well as the areas that could be related with the production of the sound.

VI. CONCLUSIONS

In this work a dataset for speaker independent lip reading with Greek words in unconstrained driving scenario is presented. This work presented a novel dataset and focused on two experiments in order to assess the ability of Deep CNNs for the task of Lip Reading (Visual Speech Recognition) in Greek words. Firstly, the ability of lip reading systems to learn speaker independent representations is evaluated by following a Leave-N persons out protocol. The results for N=2 are presented on Table III and together with the findings of the model attention indicate that these models tend to learn user dependent feature representations.

Secondly, the ability of the models to scale with respect to the number of words is evaluated. In order to investigate this, the number of words was varied in the range between 2 and 40 words, by randomly selecting words from a pool of all available 50 words. It is obvious that by increasing the number of words, the performance decreases significantly, reaching a performance plateau similar to that of human's performance. Also, by comparing human performance with the performance of an automated lip reading system, the assumption that humans categorize one word into one possible words is made. As compared to automated system, humans perform the task of lip reading in an Open-set evaluation. This means that the possible number of words spans to the user's known words. On the contrary, lip reading systems are being evaluated on a Close-set evaluation. In this case, their task is to assign a word sequence to the category (word) that mostly matches. A more robust evaluation would require having some extra words that are not learnt by the system and evaluate the automated lip reading models in both Open-set and in Close-set. In conclusion, Deep Learning architectures seem to provide a very promising solution for the task of lip reading but extra effort is required in order to allow the networks generalize well to new unseen users.

VII. Discussion

This work is expected to open new discussions in the area of automated lip reading as well as to challenge current and future research approaches in this field. Thus, from the authors perspective it is estimated that the dataset can be used as a novel benchmark for evaluating new algorithms. In this direction, the dataset is scheduled to further increased both in terms of words and persons. Additionally, the dataset will be extended in order to cover more every-day scenarios, where a lip reading usage can be developed. Regarding the research findings of this work, the authors suggest that further investigation is required in order to a) determine the limitations of these models to learn user-independent features by focusing on the CNN models that perform the feature extraction, b) investigate the applicability of Temporal Convolutional Neural Networks, as well as other forms for sequence encoding (e.g. graph based techniques) and c) explore the transferability of knowledge for the task of lip reading between different languages.

Acknowledgment

This work is partially supported by the Greek Secretariat for Research and Technology, and the EU, Project HealthSign: Analysis of Sign Language on mobile devices with focus on health services T1EΔK-01299 within the framework of "Competitiveness, Entrepreneurship and Innovation" (EPAnEK) Operational Programme 2014-2020.

References

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[2] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMS," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2592–2596.

[3] G. I. C. and, "Lipreading from color video," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1192–1195, Aug. 1997.

[4] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based Human Action Recognition via Sparse Representation in Dissimilarity Space," *J. Vis. Comun. Image Represent.*, vol. 25, no. 1, pp. 12–23, Jan. 2014.

[5] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: Sentence-level Lipreading," *CoRR*, vol. abs/1611.01599, 2016.

[6] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590–605, Sep. 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[8] J. S. Chung and A. Zisserman, "Lip Reading in the Wild," in *Asian Conference on Computer Vision*, 2016.

[9] I. Almajai, S. Cox, R. Harvey, and Y. Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2722–2726.

[10] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Continuous Automatic Speech Recognition by Lipreading," in *Motion-Based Recognition*, M. Shah and R. Jain, Eds. Dordrecht: Springer Netherlands, 1997, pp. 321–343.

[11] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "A New Visual Speech Recognition Approach for RGB-D Cameras," in *Image Analysis and Recognition*, 2014, pp. 21–28.

[12] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," *arXiv:1703.04105 [cs]*, Mar. 2017.

[13] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *CoRR*, vol. abs/1409.0473, 2015.

[15] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.

[16] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[17] D. Kastaniotis, I. Theodorakopoulos, C. Theoharatos, G. Economou, and S. Fotopoulos, "A Framework for Gait-based Recognition Using Kinect," *Pattern Recogn. Lett.*, vol. 68, no. P2, pp. 327–335, Dec. 2015.

[18] D. Kastaniotis, I. Ntinou, D. Tsourounis, G. Economou, and S. Fotopoulos, "Attention-Aware Generative Adversarial Networks (ATA-GANs)," in *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, 2018, pp. 1–5.