LRRo: A Lip Reading Data Set for the Under-resourced Romanian Language

Andrei Cosmin Jitaru
andrei.jitaru@stud.etti.upb.ro
CAMPUS-UPB
University Politehnica of Bucharest
Bucharest, Romania

Şeila Abdulamit
seila.abdulamit@stud.etti.upb.ro
CAMPUS-UPB
University Politehnica of Bucharest
Bucharest, Romania

Bogdan Ionescu
bogdan.ionescu@upb.ro
CAMPUS-UPB
University Politehnica of Bucharest
Bucharest, Romania

ABSTRACT

Automatic lip reading is a challenging and important research topic as it allows to transcript visual-only recordings of a speaker into editable text. There are many useful applications of such technology, starting from the aid of hearing impaired people, to improving general automatic speech recognition. In this paper, we introduce and release publicly lip reading resources for Romanian language. Two distinct collections are proposed: (i) wild LRRo data is designed for an Internet in-the-wild, ad-hoc scenario, coming with more than 35 different speakers, 1.1k words, a vocabulary of 21 words, and more than 20 hours; (ii) lab LRRo data, addresses a lab controlled scenario for more accurate data, coming with 19 different speakers, 6.4k words, a vocabulary of 48 words, and more than 5 hours. This is the first resource available for Romanian lip reading and would serve as a pioneering foundation for this under-resourced language. Nevertheless, given the fact that word-level models are not strongly language dependent, these resources will also contribute to the general lip-reading task via transfer learning. To provide a validation and reference for future developments, we propose two strong baselines via VGG-M and Inception-V4 state-of-the-art deep network architectures.

CCS CONCEPTS

• **Computing methodologies** → Supervised learning by classification.

KEYWORDS

visual speech recognition, lip reading, under-resourced languages, annotated data set, Romanian language

ACM Reference Format:
Andrei Cosmin Jitaru, Şeila Abdulamit, and Bogdan Ionescu. 2020. LRRo: A Lip Reading Data Set for the Under-resourced Romanian Language. In *11th ACM Multimedia Systems Conference (MMSys'20), June 8–11, 2020, Istanbul, Turkey*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3339825.3394932

1 INTRODUCTION

Lip reading, known as *Visual Speech Recognition* (VSR) from visual-only recordings of a speaker's lips, is the ability to understand or to sense the subject of the transmitted message, by a human lip reader or by a machine. Due to the complexity of task, human lip reader performances are not comparable with the machine ones, in terms of processing time and when addressing multi-language speaking prediction, the later ones being more promising. Nowadays, the progress in computer vision and deep learning offers a solution for generic lip reading via automatic deep learning techniques. The VSR has many practical applications, not only for supporting hearing impaired people, but also for transmitting instructions in noisy environments, security applications, resolving multi-talker simultaneous speech, to improving the market product recommendation and the performance of automated speech recognition, in general.

Building a VSR system requires the availability of ground truth data, i.e., explicit recordings of people pronouncing words and sentences that are annotated at utterance level. Lip reading data sets were developed to solve certain scenarios for which the VSR systems are designed. For instance, there are word-level and sentence-level data sets. For English, one of the most prominent data is the GRID corpus [2], developed for labelling computation of a particular sentence structure. Another example is the LRW [5] word-level public data set. The LRW-1000 [12] was released for Mandarin word-level lip reading. An overview of the most common data sets is presented in Table 1. Inherently, there are several limitations for each approach, some of the most important being that fact that are not being integrable with real applications (e.g., LRW [5], LRS2 [4]), having limited vocabulary and sentences (e.g., GRID [2]), or the incapacity of meeting the specific lexicon requirements of a more focused scenario, e.g., in-the-wild ad-hoc scenario (e.g., LRS3 [1]). Overall, the available data for the last four years is covering a very limited number of different languages, in particular: English (by far the most predominant), Mandarin, and Urdu.

In this paper, we propose a new resource for another under-resourced language, namely annotated data for Romanian lip reading. There are basically no other resources available so far. The only attempt towards lip reading is found in [8], where the authors analyze how accurate a 3D facial animation model for simulating visual speech production in the Romanian language should be. At least 17 words with their correspondent animation model were tested by deaf lip reading teachers.

When designing a lip reading data set, there are several specific challenges that should be addressed, which makes it a very demanding task: (i) the alignment of the audio and the spoken transcript which is necessary to generate the ground truth vocabulary or the

MMSys'20, June 8–11, 2020, Istanbul, Turkey
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6845-2/20/06…$15.00
https://doi.org/10.1145/3339825.3394932

Table 1: Overview of the existing visual speech recognition data sets (Train-Val-Test stands for training, validation and final evaluation splits, #Spk. represents the number of different speakers, #Utt. represents the number of utterances, #Words is the number of different spoken words, Vocab. represents the number of different units, e.g., words, present in the collection).

Name	Source	Split	Language	#Spk.	#Utt.	#Words	Vocab.	#Hours	Year
AVICAR	Lab	-	English	100	59k	1.3k	10	-	2004
AVLetters	Lab	-	English	10	780	26	26	-	2002
OuluVS1	Lab	-	English	20	-	817	10	-	2009
OuluVS2	Lab	-	English	52	-	9.1k	10	-	2015
GRID	Lab	-	English	34	34k	165k	51	-	2017
LRW	BBC	Train-Val	English	-	514k	514k	500	1.6k	2016
		Test			25k	25k	500		
LRS2	BBC	Pretrain	English	-	101k	4.2M	16.5k	4.9k	2017
		Train-Val			5k	358k	4.5k		
		Test			11.7k	11k	6.8k		
LRS3	TED & TEDx (YouTube)	Pretrain	English	5.5k	132k	-	52k	444	2018
		Train-Val		4k	32k		17k	30	
		Test		451	145		5.1k	1	
TCD-TIMIT	Lab	-	English	62	6.9k	-	-	-	2015
LRW-1000	Chinese TV programs	-	Mandarin	>2k	-	718k	1k	57	2018
Urdu	Lab	-	URDU	10	10	10	-	-	2018
Wild LRRo (proposed)	TV show, news program (YouTube)	Train-Val	Romanian	>35	-	1.1k	21	21	2019
		Test							
Lab LRRo (proposed)	Lab	Train-Val	Romanian	19	-	6.4k	48	5	2020
		Test							

sentences; (ii) very often, fast changes of the speaker's scene cause false positive annotations which are to be avoided; (iii) dealing with the case of multiple speakers scenes, where a system for speaker's identification is essential; (iv) a series of complementary algorithms are typically involved, e.g., a lip detector for localizing automatically the region of the mouth, automatic speech-to-text to generate the transcripts, and so on. For instance, the GRID data set [2] uses a special controlled environment for recordings, where each speaker was individually recorded with a high contrast background. There was no need for speaker identification modules. For the LRW data set [5], the authors used TV broadcasts which were processed to select the program types together with subtitle processing and aligning, face detection and tracking, speaker identification, and force aligning the transcripts with the audio signal. The authors in [12] analyze how efficient is to use voice-to-text tools along with manual annotations.

Following the best practices from the literature, we designed and released two new resources for Romanian lip reading, namely the Wild LRRo data set that features data retrieved from the Internet, i.e., in-the-wild scenario, and the Lab LRRo data set with data recorded in a controlled environment, both in terms of the target vocabulary and quality of the audio-visual information[1].

We identify the following contributions beyond state of the art: (i) this is the first resource available for Romanian lip reading and would serve as a foundation for this under-resourced language; (ii) despite the data addressing a specific language, this resource will contribute to the lip reading in general as word-level models are not strongly language dependent. They would allow for training a general system and particularizing via transfer learning; (iii) several baselines were provided for future developments, namely by training two lip reading deep architectures: MT [5] and Inception-v4 [11]. This provides also a validation of the consistency of the data and a basis reference for evaluation.

The remainder of the article is organized as follows. Section 2 introduces the proposed data sets along with their creation and annotation process. Section 3 proposes several baselines for the data. Finally, Section 4 concludes the paper and provides future perspectives.

2 PROPOSED DATA SETS

In this section, we discuss the design of the proposed data sets along with the annotations. The development process consists of a multi-stage pipeline, which interleaves manual annotation with automatic tools for processing and filtering of the data. The LRRo data contains the Wild LRRo data set and the Lab LRRo data set. These are designed such as to ensure enough samples for a lip reading learning system, e.g., sufficient number of speakers, speech rate variety and various backgrounds, targeting under-resourced languages, such as Romanian language.

The data comes with more than 1,200 minutes of TV show recordings and natural speech recordings. The data is fairly gender balanced, with 64-66% of the speakers being males. The raw videos for Wild LRRo were downloaded from YouTube[2]. The Lab LRRo was recorder in a lab controlled environment, where the speaker is in a room in front of a camera. Each instance of the LRRo data is provided as mouth crops in *.jpg* format. These were obtained

[1] to download the data, please use the following link: https://doi.org/10.5281/zenodo.3753559. For the lab recorded data, user permission was obtained and user data are anonymized. The wild data was retrieved from data that is already publicly available on the Internet.

[2] https://www.youtube.com/

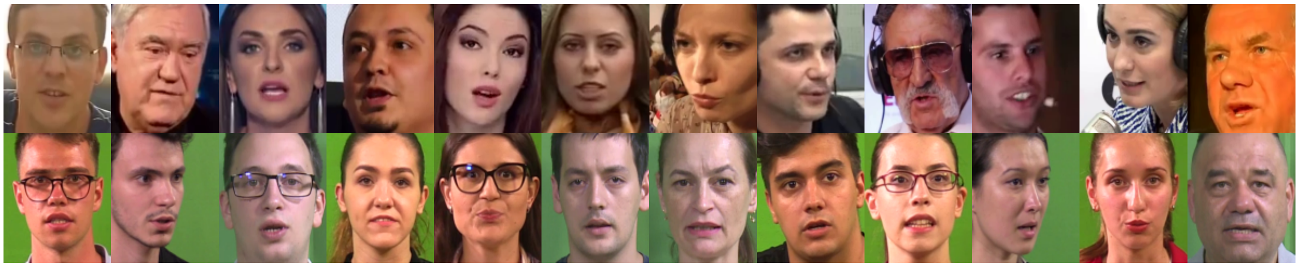

Figure 1: Sample images from the Wild (first row) and Lab (second row) LRRo data sets.

from *.mp4* short video segments which contain only relevant visual information. The segmented clips were obtained using ffmpeg[3] and the timestamps for each annotated word. Each recording was encoded using the .h264 ffmpeg codec. Several filters were applied on the segmented clips, to obtain mouth crop stacks with the same parameters. To filter the useful scenes, we have used *the useful face appearance* concept [5], i.e., a scene is relevant if a face is present continuously for 5 seconds. The face appearance of the speakers is very wide in terms of multiple-view angles, lighting conditions and make-up. Some samples are presented in Figure 1.

As a general consideration, each spoken language possesses specific linguistic elements that makes the lips movement to be challenging, from phonetic (*phoneme*) and visual (*viseme*, i.e., visual phoneme) aspects. The threshold of minimum 6 frames/instance was set to be able to detect the specific *visemes* of the Romanian language. In Romanian language, there are 20 *visemes* [8]. Excluding the cases when visual phonemes are analyzed in context dependent, according to Romanian alphabet, 4 letters with same visual information are identified (*b* and *p*; *c*, *g* and *k*; *f* and *v*; *s* and *z*). Also, in the *viseme* category should be included the Romanian phonetic groups *ce, ci, ge, gi, che, chi, ghe, ghi.* For lip reading analysis, each special group needs to be processed as an entire *viseme* sequence. Another case, which makes the Romanian lip reading task difficult to accomplish is the *x* consonant. This *viseme* can be described as two different viseme sequences, a *c* and *s* in some cases or a *g* and *z*, in others. These aspects should be taken into account when annotating the data.

2.1 Wild LRRo data set

2.1.1 Data collection. The raw videos were collected and segmented from several open source recordings of some Romanian TV shows (IT, social), TV news programmes (politic, economic and dramatic news) and Romanian TEDx talks. 52 large videos were retrieved. The resolution of the videos varies between 360x450 pixels and 720x1280 pixels. The videos were recorded at 25 fps. All the videos were transcribed manually by several annotators. More than 1,200 minutes of recordings were processed and segmented.

The data is challenging due to the large variations in the speech conditions, including lighting conditions, variations in pose, multi-person scenes and various emotions of the speaker, e.g., shame, anger or joy, speech rate, age, gender and make-up, the presence of interrupted shot frames.

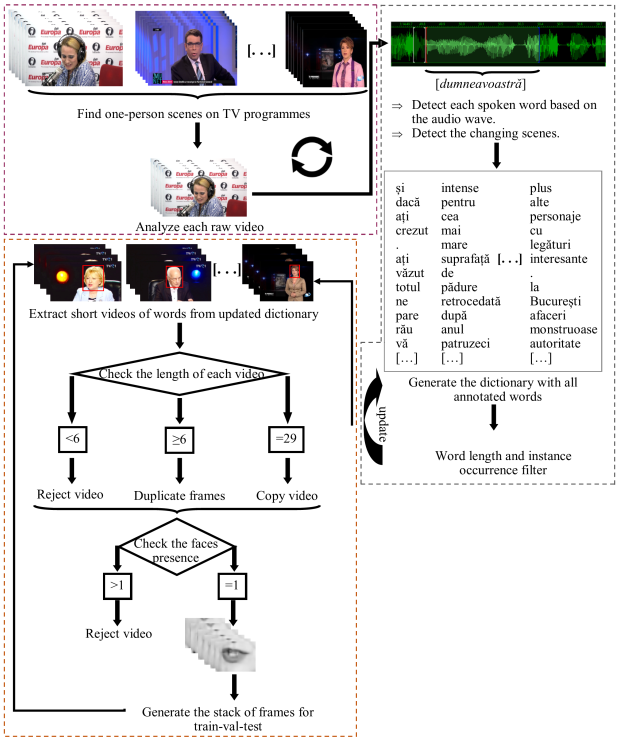
[. . .]
Find one-person scenes on TV programmes
Analyze each raw video
[dumneavoastră]
⇒ Detect each spoken word based on the audio wave.
⇒ Detect the changing scenes.
și dacă ați crezut . ați văzut totul ne pare rău vă […]
intense pentru cea mai mare suprafață de pădure retrocedată după anul patruzeci […]
[. . .]
plus alte personaje cu legături interesante . la București afaceri monstruoase autoritate […]
Generate the dictionary with all annotated words
update
Word length and instance occurrence filter
[. . .]
Extract short videos of words from updated dictionary
Check the length of each video
<6
≥6
=29
Reject video
Duplicate frames
Copy video
Check the faces presence
>1
=1
Reject video
Generate the stack of frames for train-val-test

Figure 2: Annotation process for the Wild LRRo data set.

2.1.2 Data annotation. The audio spectrogram was used to ease the annotation process by making use of the articulation between spoken words. Annotation is carried out using the Aegisub[4] tool. 18 voluntary annotators were double checked by 2 expert master annotators to provide text transcripts of every word, including the metadata. Based on their input, the raw videos are divided into individual sentences/phrases using the punctuation in the transcripts. The sentences are separated by full stops. Each spoken word with metadata is stored in a *.txt* file. The metadata consists of the start-end timestamps of the annotated word, ID of the speaker and gender (for statistical purpose). The annotation process is depicted in Figure 2.

The obtained annotations were used to further split the raw videos into multiple clips, each one corresponding to a word. Clips

[3] https://www.ffmpeg.org/

[4] http://www.aegisub.org/

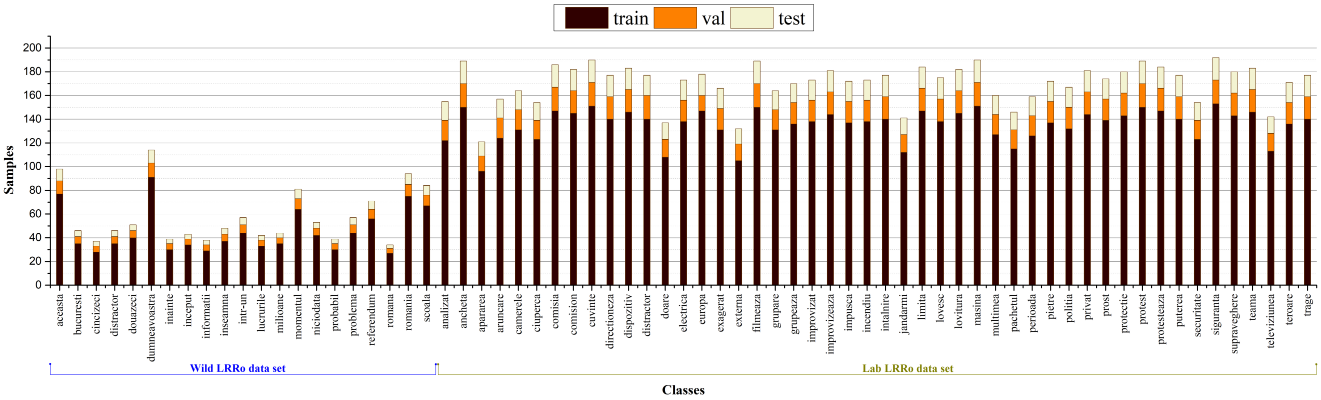

Figure 3: LRRo data sets train-val-test split.

of same word are not limited to a specified length to allow the existence of various speech rates. An intra-class slight variation of 11-13 and 12-18 frames was noticed, with no outliers. The median line of the most variable class is around 16 frames. Due to large variation of the length for each instance, the last frame of each instance was duplicated to achieve 29 frames/instance (which fits the longest word from data set). The difference in length between instances of the same class, emphasize large variations of the speech rate in the data.

Based on the annotations, metadata and the list of annotated words, the dictionary is obtained by selecting most 20 occurring words with at least 6 characters. Given the fact that the videos were not recorded in a lab environment, controlling the occurrence of the words, i.e., to be represented with enough samples, is impossible. The rest of the annotated words were grouped to form a distractor class. A class consists of several instances of a specific annotated word. Taking into account the distractor, the data comes with 21 (unbalanced) classes. The distractor class was under-sampled to improve the accuracy of the systems [13].

The final lexicon of the Wild LRRo data set consists of the following classes: *"aceasta", "bucuresti", "cincizeci", "douazeci", "dumneavoastra", "inainte", "inceput", "informatii", "inseamna", "intr-un", "lucrurile", "milioane", "momentul", "niciodata", "probabil", "problema", "referendum", "romana", "romania", "scoala"* plus the distractor class. As we previously mentioned, this data is inherently unbalanced.

2.2 Lab LRRo data set

2.2.1 Data collection. Simultaneous recordings were achieved by frontal and 30° left positioning of a second camera along with the head position of the speaker. This was done to provide some variation of viewing angle to get closer to an unconstrained scenario. We used two SONY PWN-EX1 cameras and two SONY PWM-EX3 cameras which record MPEG-2 videos. Each camera was set up to capture full frames at 25fps. The audio recordings were performed using an omni-directional lavaliere microphone, connected to transmitter-receiver pair, SONY UTX-B2 and SONY URX-P2. The trials were performed in an acoustically isolated booth. These setups ensured high quality uncompressed audio recording at 16-bit and 48kHz.

19 speakers (mostly students) were requested to read a predefined text from a prompter. Some of them have specific reading or speaking difficulties, like dyslectic or rhotacism. The speed of the prompter was adjusted to fit to their speech rate. The frontal camera was integrated into the prompter assembly. The speakers were selected thus to achieve gender balance. The predefined text was conceived thus to ensure enough representativity for a target vocabulary of words. 74 sentences were designed to make reading easier of all words from the preset lexicon.

2.2.2 Data annotation. It builds on the pipeline used for the wild data set. A transcript was generated with fixed number of verbs, adjectives and nouns, to develop a more task focused vocabulary data set. From the recording sessions resulted 27 raw videos, which contain more than 6k spoken words. In the process of annotation, all 27 recordings were manually processed to cut out the scenes without relevant information, like continuous speech or head positions that wouldn't allow tracking the lips. The synchronization of the video and the audio signal did not changed during the generation of the processed videos. Frame differences between different speakers is constant even for short words. Considering the fact that all speakers read the same words, the variation of the speech rates between similar classes is almost similar, i.e., a 4-5 frames variation was noticed.

The annotation of the videos was achieved via two different ASR (automatic speech recognition) systems, proposed in [6], and in [7]. We have compared the transcription results with our generated speech text (containing specific lexicon). The speech-to-text models were trained on an in-house developed speech Romanian corpus. Based on the ASR system's output (timestamps and annotated word), more than 6k segmented clips were obtained using ffmpeg. For a series of generated sequences, using the Aegisub software, we analyzed the location of the spoken word with respect to the visual and audio information. Afterwards, the sequences were passed through the pipeline presented in Figure 2.

The lexicon of the Lab LRRo data set consists of the following classes: *"analizat", "ancheta", "apararea", "aruncare", "camerele", "ciuperca", "comisia", "comision", "cuvinte", "directioneaza", "dispozitiv", "doare", "electrica", "europa", "exagerat", "externa", "filmeaza",* *"grupare", "grupeaza", "improvizat", "improvizeaza", "impusca", "incendiu", "intalnire", "jandarmi", "limita", "lovesc", "lovitura", "masina", "multimea", "pachetul", "perioada", "pietre", "politia", "privat", "prost", "protectie", "protest", "protesteaza", "puterea", "securitate", "siguranta", "supraveghere", "teama", "televiziunea", "teroare", "trage"* plus a distractor class with annotated words with less than 6 characters.

2.3 Data distribution

Some basic statistics about the data are presented in Table 1. The Wild LRRo data comes with more than 35 different speakers, 1.1k words out of which a vocabulary of 21 different words was possible, i.e., selection of words with a frequency of appearance enough for a learning system. The total duration of the source data is around 21 hours. The Lab LRRo data comes with 19 different speakers, 6.4k words and a vocabulary of 48 words. The total duration of the source data is of 5 hours. Although these data may seem limited compared to some more consecrated data sets, they represent a pioneering data for the Romanian language and provide enough information to train a deep learning system, as the results show (see Section 3).

The LRRo data is distributed into three subsets, namely a training subset (train) intended for training the models, a validation subset for validating and optimizing methods' parameters (val), and a final testing subset (test) for the actual evaluation. The recommended partitions are: (i) for Wild LRRo: train 846 samples, val 120 samples, and test 121 samples; (ii) for Lab LRRo: train 6,505 samples, val 860 samples, and test 815 samples. The exact distribution of the word classes within the subsets is presented in Figure 3.

To facilitate accessing easily the data for building the machine learning models, the information is structured as presented in Figure 4. Each data set has its own folder which contains a separated sub-folder for each train-val-test subsets. Then, each word has its own folder (the name of the folder is basically the ground truth) and a list of sub-folders for each of the speakers (speaker's id is stored with the folder name). Visual information is provided via the video frames stored in *.jpg* format.

3 BASELINE SYSTEMS

Lip reading prediction is an inherently multi-class learning process. In our case, each different spoken word corresponds to an individual class. The expected output of a classifier consists of the label of the predicted word class or top-k labels of the most plausible spoken words. In this section, we provide two *baseline systems* by experimenting with some popular deep neural networks architectures. Considering the performances of the VGG-M architecture [3] and Inception-V4 network [11] on various classification tasks, we select them as baselines for our data (VGG-M is implemented via the MT architecture [5]). The challenge of our systems is represented by how two lip reading models react on the frequently co-articulations which occur in the unconstrained (wild) or lab conditions, considering all the similar visual phonemes specific to Romanian language (see Section 2).

3.1 Parameter tuning

The MT architecture "ingests" 29 input frames, particularly based on the VGG-M model, which has good capabilities for distinguishing between several classes. Each of the 29 frames is taken as input by a convolutional layer with shared weights. The activations from the towers are concatenated after a pooling layer, producing an output activation with 1,344 channels. Futher, a conv1D is applied to reduce the number of activations for the second convolutional layer. The remainder of the network is the same as for the regular VGG-M. The MT network is trained using SGD with momentum [10] 0.8 and batch normalization [9], but without dropout.

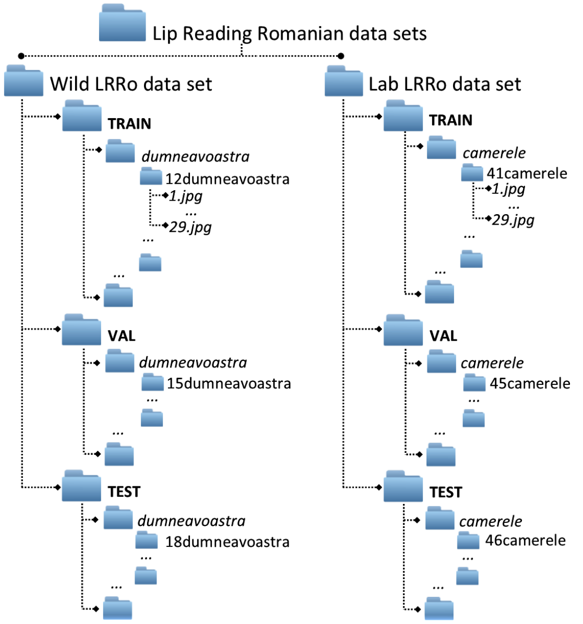

Figure 4: LRRo data structure.

In the same manner, we have trained the Inception-V4 model, described in [11]. This model keeps the core of the regular Inception model and has been optimized in terms of accuracy and training speed on ImageNet. Each input frame passes through the same layers in the "stem" block of the model, due to the shape of each sample $(64 \times 1, 856)$. Then, the weights are shared and concatenated. The output, consisting of the weights of each 64×64 frame that are loaded into the Inception blocks. Compared with the MT architecture, Inception-v4 is a deeper and highly customizable neural network. For training we use the ADAM optimizer with default parameters.

For both models, the starting learning rate was 2×10^{-4}, decreasing it with the involution of the validation accuracy. The generalisation level of the models was evaluated using the top-k accuracy which determines the total number of instances that are correctly predicted from the top ranked k.

3.2 Results

For the training of the models, we use the train-val-test splits as presented in Section 2.3. The best results for Romanian lip reading models are achieved in the case of the Lab LRRo data set, regardless of the number of classes in the training subset, which is more or less expected, as this data is fully controlled. The Wild LRRo data set is specific to Internet data with high variation of speech rate,

Table 2: Performance of the baseline systems trained on LRRo data sets.

Accuracy Data set structure		MT				Inception-V4			
		Test		Train-val		Test		Train-val	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Wild LRRo data set	21 classes	33%	61%	37%	68%	33%	62%	40%	64%
Lab LRRo data set	16 classes	76%	97%	80%	97%	75%	97%	80%	98%
	32 classes	81%	95%	82%	95%	77%	96%	81%	96%
	48 classes	71%	90%	71%	91%	71%	92%	71%	93%

co-articulations, dialect of the speaker, and dynamic behaviour of the subjects during their speech, being very difficult to pinpoint locate each word in the speech. The best results achieved with the baselines on the test data are: (i) for Wild LRRo, best Top-1 accuracy of 33% and Top-5 accuracy of 62%; (ii) for Lab LRRo, best Top-1 accuracy of 71%, and Top-5 accuracy of 92%.

Analyzing the results, we noticed that the MT architecture offers good accuracy on classes with same *visemes* (see Section 2) and for derived words as "*comisia*", "*comision*" or "*lovesc*", and "*lovitura*". The Lab LRRo training subsets are designed with alphabetically ordered classes. This is the reason why in each training subset there are pairs of derived words. The effect of these pairs is visible when the entire Lab data set was used for training the models. An almost constant generalization level of the models is described by the slightly differences of accuracy evaluation metric, as seen in Table 2. It is important to mention that the balance of the number of samples from each class and the number of samples in general of each data set has a direct influence on the accuracy.

4 CONCLUSIONS

We proposed a publicly available data for (under resourced) Romanian automatic lip-reading, namely the LRRo data: wild LRRo for an ad-hoc, Internet-like scenario and lab LRRo for a lab controlled environment and higher quality. To prepare a consistent data set, we have filtered the instances stage-by-stage, taking into account the word lengths, the number of faces in the scene, and a minimum duration of a spoken word for the proposed lexicon. The data comes with word-level trusted annotations obtained in a semi-automatic manner and curated manually by experts. This is the first resource available for the Romanian language. Nevertheless, the data is not restricted to this language as transfer learning can be used for exploiting it in other cases. Many similar examples were successfully experimented in the literature. To serve as a strong baseline, we proposed two deep network architectures (VGG-M and Inception-V4) that were experimented on the released data. Results show the potential of these resources. Future work mainly consists of extending the data with more annotated resources. A good lead is to exploit the generative power of GAN networks. Successful experiments were already conducted for generating realistic faces and therefore there is a good potential of extending them to lips.

ACKNOWLEDGEMENTS

Andrei Cosmin Jitaru and Bogdan Ionescu work was supported by the Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002. Șeila Abdulamit work has been funded by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125. The authors would like to thank Horia Cucu and Alexandru Lucian Georgescu from SpeeD lab, UPB, for providing the Romanian ASR transcripts.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *ArXiv* abs/1809.00496 (2018).

[2] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. 2016. LipNet: Sentence-level Lipreading. *ArXiv* abs/1611.01599 (2016).

[3] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets. *ArXiv* abs/1405.3531 (2014).

[4] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2016. Lip Reading Sentences in the Wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 3444–3453.

[5] Joon Son Chung and Andrew Zisserman. 2016. Lip Reading in the Wild. In *ACCV*.

[6] Alexandru-Lucian Georgescu, Horia Cucu, and Corneliu Burileanu. 2017. SpeeD's DNN approach to Romanian speech recognition. *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (2017), 1–8.

[7] Alexandru-Lucian Georgescu, Horia Cucu, and Corneliu Burileanu. 2019. Kaldi-based DNN Architectures for Speech Recognition in Romanian. *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (2019), 1–6.

[8] Mihai Daniel Ilie, Cristian Negrescu, and Dumitru Stanomir. 2013. An efficient 3 D Visual Speech Synthesis Framework for Romanian Language Logopedics use. *Romanian Journal of Information Science and Technology (ROMJIST)* (2013).

[9] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv* abs/1502.03167 (2015).

[10] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. 2013. On the importance of initialization and momentum in deep learning. In *ICML*.

[11] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI*.

[12] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. 2018. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)* (2018), 1–8.

[13] Show-Jane Yen and Yue-Shi Lee. 2006. Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset. (2006), 731–740. https://doi.org/10.1007/978-3-540-37256-1_89