IberSPEECH 2022
14-16 November 2022, Granada, Spain

Spanish Lipreading in Realistic Scenarios: the LLEER project

Carlos-D. Martínez-Hinarejos, David Gimeno-Gómez, Francisco Casacuberta, Emilio Granell, Roberto Paredes, Moisés Pastor-Gadea, Enrique Vidal

Pattern Recognition and Human Language Technologies Research Center
Universitat Politècnica de València, Camí de Vera, s/n, 46022, València, Spain
{cmartine,dagigo1,fcn,egranell,rparedes,mpastorg,evidal}@dsic.upv.es

Abstract

Automatic speech recognition has been usually performed by using only the audio data, but speech communication is affected as well by other non-audio sources, mainly visual cues. Visual information includes body expression, face expression, and lip movements, among other. Lip reading, also known as Visual Speech Recognition, aims at decoding speech by only using the image of the lip movements. Current approaches for automatic lip reading follow the same lines than for speech processing: use of massive data for training deep learning models that allow to perform speech recognition. However, most of the datasets and models are devoted to languages such as English or Chinese, while other languages, particularly Spanish, are underrepresented. The LLEER (*Lectura de Labios en Español en Escenarios Realistas*) project aims at the acquisition of large-scale visual corpora for Spanish lip reading, the development of visual processing techniques that allow to extract important information for the task, the implementation of models for automatic lip reading, and the integration with speech recognition models for audiovisual speech recognition.

Index Terms: lip reading, audiovisual speech recognition, image and video processing

1. Introduction and background

Speech recognition is a classic problem in the computer science field [1] which can be considered as a complex problem where many different senses are involved, as well as high-level knowledge processes such as grammar, semantics and pragmatics [2]. Most of the work on automatic speech recognition is based on the use of audio signal (voice), but other information sources can be useful: gestures, face expressions, body pose, etc. Actually, speech production involves a mechanical process where lip movements are involved. Some studies [3] concluded that coherence between lip movements and audio signal is fundamental for speech intelligibility. Therefore, lip movements could be of great help to improve speech decoding with respect to using only audio signal (audiovisual speech processing [2, 4]), which is useful specially when audio signal is noisy or corrupted, while video signal presents a relatively better quality [4].

A step forward would be the use of only the visual modality for speech decoding, which is known as automatic lip reading [5] or Visual Speech Recognition (VSR). This approach is quite more difficult because only about 30% of speech information is visible [6]. Having an automatic system that would be able to decode speech from lip movements, would be a valuous tool for people with that functional diversity (i.e., people with difficulties on producing audio signals).

Several factors have influenced automatic lip reading in its development in the last decades. Initial attemps of VSR were related to very simple identification tasks such as alphabet and digit recognition [7], based on corpora such like AVLetters [8] and CUAVE [9]. Steps towards continuous VSR were taken for a single speaker [10], using the RM-3000 corpus and for several speakers [11] with the TCD-TIMIT corpus. However, these corpora were recorded in controlled conditions, which makes them not fully useful for working in a real scenario. A more realistic environment is that compiled in the LRS2 corpus [12], a set of videos taken from BBC broadcasts, and the LRS3 corpus [13], related to TEDx talks. All these corpora are available only for the English language.

Regarding to other languages, and specially Spanish language, the number of datasets is really lower. A popular one is the VLRF corpus [14], which is a freely available corpus of visual recordings in Spanish in controlled conditions. This corpus allowed to obtain a lip reading system with modest results [14]. The Av@Car corpus [15] is another example of Spanish dataset that can be used for lip reading, in this case directed towards audiovisual speech recognition in a noisy environment (car driving). Another dataset is the CMU-MOSEAS corpus [16], which presents video samples in unrestricted conditions in 4 different languages, including Spanish.

Apart from data availability another principal aspect for VSR is the feature extraction step. Feature extraction techniques are derived from usual image and video processing techniques [17], such as the use of geometric features [18], Eigenlips [19], and Deep Learning approaches based on Convolutional Neural Networks (CNN) [20], and more specifically Convolutional Autoencoders [21] and the VGG model [22]. At the moment, it seems that there is no clear alternative that works better than other for the lip reading task [7].

VSR is nowadays mainly based on the Machine Learning paradigm [23] with models that have evolved in a similar manner to that of ASR models. Hidden Markov Models (HMM) with Gaussian Mixtures (GMM-HMM) was the usual approximation since a decade ago [24]. Nowadays, most models are based on Deep Learning, initially hybrid HMM (DNN-HMM) [25, 26] and nowadays End-to-End models [12], in which input video frames are directly decoded into the word sequence. In any case, current state-of-the-art models provide a promising but still high error (around 30%) for English VSR [27, 28, 29], which reflects the intrinsic difficulty of the task. In Spanish, this situation is even worse due to the lack of resources, having around a 50% of error rate in different datasets [28].

In this context, the LLEER project aims at contributing to the advance of the lip reading technology for the Spanish language and its application in unrestricted scenarios. In contrast to other proposals, LLEER would provide a large scale visual corpus for Spanish lip reading in unrestricted conditions. This corpus would be based on the Albayzín RTVE Challenge

10.21437/IberSPEECH.2022-49

dataset [30], which provides quite realistic scenarios with an unrestricted nature of the images, because it contains TV programs from *Radio Televisión Española* (RTVE) that cover from news to magazines, and in different scenarios (TV studios, interviews or shows in common places, street scenes, etc.). Part of this corpus is available as a set of videos with their corresponding transcriptions, and it would be the basis for generating a large visual corpus of lip movements with their transcriptions. Since some parts contain no lip image available (off-voice, speaker not focused by the camera, speaker not talking to the camera, ...), it must be properly processed to obtain the final dataset that could be used for developing the rest of the activities proposed in the project. This corpus would be made freely available (considering the conditions imposed by the RTVE corporation) in order to provide a valuable resource for the research community on VSR, and specially to those working on the Spanish language.

The project would also study different feature extraction techniques in order to determine which is the best general approach for feature extraction. The obtained conclusions would guide future research in this field by showing which is the most convenient feature extraction technique given the recording conditions.

Additionally, extensive experimentation with different Machine Learning models is planned. At the initial steps of the project, a limited amount of data would be available (those freely available datasets and a part of the RTVE dataset) and possibly not in the desired conditions (unrestricted scenarios). Thus, initial approaches based on classic models would be experimented, to switch to state-of-the-art models (those based on Deep Learning and End-to-End models) as more data becomes available. The experience on the previous realisations would allow to make a more efficient use of the available resources (data, computational resources, human resources, etc.). The final result would be a model architecture that could provide state-of-the-art results for lip reading in Spanish, and that would be the basis for future works by other researchers in the area.

Finally, the integration of the developed lip reading module into an audiovisual recognition system would be performed, along with its implementation in a software tool that would allow to perform automatic lip reading (or audiovisual recognition) for people that are interested in their use, specially for people with functional diversity that would improve their life quality by using that tool.

2. Objectives and tasks

The main objectives of the LLEER project are:

1. To compile a corpus for Spanish VSR in unrestricted conditions.
2. To develop feature extraction techniques for VSR.
3. To create models that perform automatic VSR.
4. To integrate VSR with audio recognition to implement Audio Visual Speech Recognition (AVSR).
5. To implement a software tool for VSR/AVSR.

To obtain these objectives, the project is divided into work packages (WP) and tasks, which are presented in the rest of this section.

2.1. WP1. Corpus acquisition and organisation

This WP is related to the obtention of the corpora that would be used for feature extraction research and for training and testing the models to be developed. There are different available datasets with different features in Spanish and other languages to be considered. Apart from that, a fully new corpus based on the Albayzín RTVE Challenge corpus would be built in order to obtain a dataset with enough data in realistic conditions.

The tasks are the following:

1. Dataset selection: there are available a large set of audiovisual datasets that are related to the lip reading task, each one with a different language, degree of complexity, and recording conditions. This task would perform an exploration of the different corpora and their availability. The RTVE corpus will be in the set of selected corpora.
2. Dataset compilation: the datasets would be selected according their features and the capacities to process them. Datasets would be downloaded (video and transcriptions) along with their corresponding standard training/validation/test partitions (if available), and stored in a centralised server.
3. Dataset processing and annotation: since most datasets are not fully suitable for the lip reading task (such as the RTVE dataset), a processing of the corpus would be necessary to select only the material that is suitable for lip reading. This can be performed semi-automatically, detecting video frames where faces are present and then performing a manual revision to disregard unsuitable material (silent people, faces that are not looking at the camera, frames with many faces, etc.) and build the final set. This set of sequences must be properly annotated from the provided transcription.
4. Dataset organisation: datasets should be properly partitioned in training, validation and test parts. Many different partitions are possible: speaker dependent or independent, closed or open language model, similar or different recording conditions, etc. A meta-corpus could be organised with material from different corpora in order to achieve a larger set of sequences in some given conditions. This task would define experimental frameworks and its documentation. Possible evaluation campaings could be proposed in the framework of the Albayzín evaluations performed every two years in the framework of the IberSpeech conference.

2.2. WP2. Feature extraction

This WP is related to the extraction of the most relevant features for the datasets compiled and selected in WP1. Different feature extraction techniques would be studied, selected, and proposed. Appropriate feature combination techniques would be implemented as well in order to obtain a set of features that combines the best characteristics of each technique.

The related tasks are the following:

1. Study and selection of feature extraction techniques: an initial study of the proposed feature extraction techniques would be performed, in order to deepen in the characteristics of each technique, their problems and limitations, and their suitability for different scenarios.
2. Proposal and implementation of feature extraction techniques: the selected techniques would be implemented and applied to the available corpora. Initial results would allow to have a better knowledge of the limitations of each approach and to change the different parameters in the extraction steps for optimal performance.

3. Feature combination: the use of feature combination techniques has provided good results in different recognition and classification tasks. Therefore, the features selected and implemented would be combined and tested with different feature combination techniques.

2.3. WP3. HMM-based decoders

Although nowadays Deep Learning systems without the use of HMM are the common option, HMM combined with Deep Neural Networks (DNN) are still a state-of-the-art solution for speech recognition and VSR. The objective in this work package is to implement a final DNN-HMM system from the selected datasets that could obtain enough precission.

The proposed tasks are:

1. Software selection: there are different software options, being Kaldi one of the most popular; it could be complemented with other tools for managing the different parts of the system (lexical construction, language models, lattice processing, etc.).
2. GMM-HMM training and testing: a base GMM-HMM model is usually necessary for intialising a DNN-HMM model. In this task it would be generated based on the selected datasets and on different morphological units (phonemes or visemes).
3. DNN-HMM training and testing: by using GMM-HMM models, DNN-HMM models could be initialised and trained with the defined datasets. Different setups would be experimented, and data augmentation techniques would be used to obtain a higher amount of data for better training. Techniques such as Sequence Discriminative Training would be applied.

2.4. WP4. End-to-End decoders

Current state-of-the-art architectures for VSR are mainly based on Deep Learning. The visual nature of the VSR task makes it suitable for its processing by Convolutional Neural Networks (CNN) and the time series nature of VSR fits the paradigm of Recurrent Neural Networks (RNN). All these architectures could be integrated in an End-to-End fashion that, currently, provides state-of-the-art results for corpora in unrestricted conditions. This work package will study and propose different architectures for Spanish VSR in unrestricted conditions.

The proposed tasks are:

1. Review of Deep Learning methods for VSR: current state-of-the-art of Deep Learning methods for VSR is evolving rapidly, with publications in document repositories without peer-review processing but with valuable contributions. Thus, it would be necessary to review periodically what are the trends in Deep Learning for this task, with an special emphasis in techniques centered in the End-to-End paradigm.
2. Selection of software tools: the implementation of Deep Learning models is supported by several software libraries, such as Keras, TensorFlow, Caffe, or PyTorch, with different features and available resources. According to our needings in the implementation of the models for VSR, one or several of these tools would be selected for the different implementation steps.
3. End-to-end models implementation, training, and test: the defined architectures would be implemented, trained, and tested according to the data selection provided in WP1. Different strategies such as Transfer Learning or data augmentation would be used to improve performance. Transfer Learning could be of particular interest, since there are many pretrained models specialised on face image processing.

2.5. WP5. Audiovisual integration

The best VSR option could be integrated within an Audiovisual Speech Recognition (AVSR) system that could take advantange of the different error sources of each modality in order to obtain a better result. There are several integrating options, such as feature integration or hypothesis integration. As a result, a final AVSR system would be produced.

The tasks associated are the following:

1. Audio-based speech recognition system: a base system for each modality is necessary, such that their outputs can be processed. Therefore, a basic audio-based speech recognition system would be implemented.
2. Feature integration AVSR model and system: AVSR could be obtained from feature combination. Several works that cover this approximation and their proposals would be reviewed and selected.
3. Hypothesis integration AVSR system: another option for building an AVSR system would consist of decoding each modality by separate and then combining the produced outputs. This can be done by obtaining output lattices for each modality and performing a fusion of the two lattices. Several fusion techniques are available and they will be explored and implemented.

2.6. WP6. VSR/AVSR recognition tool

All the work in the project would be reflected in a software tool that would be able to decode visual and audio-visual speech. The software tool would have access to the proper devices (camera and microphone) to make the acquisition of the video and audio modalities and would provide them to the recognition system in order to obtain its decoding and presenting it to the user.

The tasks related to this objective are:

1. Acquisition software: standard libraries would be explored for the acquisition of video and audio in a computer system and in mobile devices. A tool with a simple interface that allows for video and audio acquisition and transmision would be developed.
2. Decoder integration: VSR and AVSR are computationally demanding tasks; therefore, a client-server architecture would be proposed, in which the client performs the acquisition and it sends to the server the acquired signals. The decoding mode would be selected from the client and the server would chose the correct system (visual or audio-visual), according to what client sent, to perform the recognition, whose result would be sent back to the client.

3. Current achievements

Although the LLEER project officially started in September 1st 2022, previous work on their topics has been performed and several tasks and objectives have been partially covered. In particular:

- Corpora selected: The exploration of available corpora allowed us to obtain different datasets with different features; more specifically, for Spanish we obtain the datasets VLRF [14], CMU-MOSEAS [16], and Multilingual-TEDx [31] (these two last corpora contain data for other languages as well); for transfer learning purposes we selected other English corpora, such as LRS2 [12] and LRS3 [13].
- Corpora acquired: the LIP-RTVE corpus was edited and annotated properly from a subset of videos of the Albayzín RTVE corpus. The corpus is now publicly available[1] and it has be publicly presented in the LREC conference [32].
- DNN-HMM systems: HMM-based systems have been created for the VLRF and the LIP-RTVE datasets; GMM-HMM results have been provided as baseline results for LIP-RTVE in [32], and DNN-HMM results have been obtained for VLRF and LIP-RTVE which provided some improvements. In the case of VLRF, a technique based on previous audio alignments was used to obtain the best results so far in this corpus using HMM-based models.
- End-to-End systems: based the recent proposal presented in [28], several End-to-End systems have been implemented and tested with both English datasets (mostly for initialising and perform transfer learning) and Spanish datasets such as LIP-RTVE, CMU-MOSEAS, and Multilingual-TED. The obtained results are promising and showed a large improvement with respect to using HMM-based systems, as it could be expected.

4. Conclusions and future work

This article presents the objectives and tasks of the LLEER project, which it is aimed at developing VSR systems for Spanish that could be integrated into AVSR systems and in a software tool that could be useful for people with speech production problems or for other applications (such as silent passwords, i.e., speech passwords where no audio is uttered but only lip movements are available). Although the project has been recently started, some previous work allowed to obtain in a short time many different results, including an initial version of the LIP-RTVE database and some lip reading systems that are providing promising results.

Future work is directed towards the research of the rest of objectives, such as the increase of the number of datasets in Spanish to be used, the extension of the LIP-RTVE dataset, the exploration of different feature extraction techniques and in different acquisition conditions (light variation, distance of camera to the speaker, ...), the proposal of new End-to-End architectures for VSR, the integration of the created systems into AVSR systems and the development of a software application, preferably a mobile device app, that performs Spanish lip reading with accurate enough results.

5. Acknowledgements

Work partially supported by project LLEER, PID2021-124719OB-I00 funded by MCIN/AEI/10.13039/501100011033/ ERDF, EU.

[1] https://github.com/david-gimeno/LIP-RTVE

6. References

[1] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, ser. Signals and Communication Technology. London: Springer, 2015.

[2] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976.

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

[5] I. Rodríguez Ortiz, "Lipreading in the prelingually deaf: what makes a skilled speechreader?" *The Spanish journal of psychology*, vol. 11, no. 2, pp. 488–502, 2008.

[6] P. Duchnowski, D. Lum, J. Krause, M. Sexton, M. Bratakos, and L. Braida, "Development of speechreading supplements based on automatic speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 4, pp. 487–496, 2000.

[7] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing*, vol. 78, pp. 53–72, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885618301276

[8] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.

[9] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2002, pp. II–2017–II–2020.

[10] D. Howell, S. Cox, and B. Theobald, "Visual units and confusion modelling for automatic lip-reading," *Image and Vision Computing*, vol. 51, pp. 1–12, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885616300294

[11] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[12] J. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 3444–3453. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.367

[13] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," 2018. [Online]. Available: https://arxiv.org/abs/1809.00496

[14] A. Fernandez-Lopez, O. Martinez, and F. M. Sukno, "Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2017, pp. 208–215. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/FG.2017.34

[15] A. Ortega, F. Sukno, E. LLeida, A. Frangi, A. Miguel, L. Buera, and E. Zacur, "AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/389.pdf

[16] A. Bagher Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, "CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP). Online: Association for Computational Linguistics, Nov. 2020, pp. 1801–1812. [Online]. Available: https://aclanthology.org/2020.emnlp-main.141

[17] A. Bovik, *Handbook of Image and Video Processing*, ser. Communications, Networking and Multimedia. Elsevier Science, 2010. [Online]. Available: https://books.google.es/books?id=UM_GCfJe88sC

[18] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015, pose & Gesture. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314215002088

[19] K. Thangthai, H. L. Bear, and R. Harvey, "Comparing phonemes and visemes with dnn-based lipreading," 2018. [Online]. Available: https://arxiv.org/abs/1805.02924

[20] S. Khan, H. Rahmani, S. Shah, M. Bennamoun, G. Medioni, and S. Dickinson, *A Guide to Convolutional Neural Networks for Computer Vision*, ser. Synthesis Lectures on Computer Vision. Morgan & Claypool Publishers, 2018. [Online]. Available: https://books.google.es/books?id=e0NRDwAAQBAJ

[21] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[23] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[24] M. Zimmerman, "Visual speech recognition: from traditional to deep learning frameworks," Ph.D. dissertation, 2018.

[25] K. Thangthai, R. Harvey, S. J. Cox, and B.-J. Theobald, "Improving lip-reading performance for robust audiovisual speech recognition using dnns," in *AVSP*, 2015, pp. 127–131.

[26] K. Thangthai and R. Harvey, "Improving computer lipreading via dnn sequence discriminative training techniques," in *Proceedings of Interspeech 2017*. ISCA, Aug. 2017, pp. 3657–3661.

[27] K. R. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," *arXiv preprint arXiv:2110.07603*, 2021.

[28] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," 2022. [Online]. Available: https://arxiv.org/abs/2202.13084

[29] B. Shi, W. N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[30] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Rtve2020 database description," 2020. [Online]. Available: catedrartve.unizar.es/reto2020/RTVE2020DB.pdf

[31] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "The Multilingual TEDx Corpus for Speech Recognition and Translation," in *Proc. Interspeech 2021*, 2021, pp. 3655–3659.

[32] D. Gimeno-Gómez and C.-D. Martínez-Hinarejos, "LIP-RTVE: An audiovisual database for continuous Spanish in the wild," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 2750–2758. [Online]. Available: https://aclanthology.org/2022.lrec-1.294