IberSPEECH 2022
14-16 November 2022, Granada, Spain

Speaker-Adapted End-to-End Visual Speech Recognition for Continuous Spanish

David Gimeno-Gómez , Carlos-D. Martínez-Hinarejos

Pattern Recognition and Human Language Technologies Research Center,
Universitat Politècnica de València, Camino de Vera, s/n, 46022, València, Spain
dagigo1@dsic.upv.es, cmartine@dsic.upv.es

Abstract

Different studies have shown the importance of visual cues throughout the speech perception process. In fact, the development of audiovisual approaches has led to advances in the field of speech technologies. However, although noticeable results have recently been achieved, visual speech recognition remains an open research problem. It is a task in which, by dispensing with the auditory sense, challenges such as visual ambiguities and the complexity of modeling silence must be faced. Nonetheless, some of these challenges can be alleviated when the problem is approached from a speaker-dependent perspective. Thus, this paper studies, using the Spanish LIP-RTVE database, how the estimation of specialized end-to-end systems for a specific person could affect the quality of speech recognition. First, different adaptation strategies based on the fine-tuning technique were proposed. Then, a pre-trained CTC/Attention architecture was used as a baseline throughout our experiments. Our findings showed that a two-step fine-tuning process, where the VSR system is first adapted to the task domain, provided significant improvements when the speaker adaptation was addressed. Furthermore, results comparable to the current state of the art were reached even when only a limited amount of data was available.

Index Terms: Visual Speech Recognition, Speaker Adaptation, Spanish Language, End-to-End Architectures

1. Introduction

Influenced by studies which have demonstrated the relevance of visual cues throughout the speech perception process [1] , different advances have been achieved in the field of Speech Technologies. The robustness of automatic speech recognition systems, specially in adverse conditions where the acoustic signal was damaged or corrupted [2], has been enhanced by the design of audio-visual approaches [3, 4]. Moreover, these studies have encouraged the development of systems capable of interpreting speech by reading only the lips of the speaker. In fact, this challenging task, known as Visual Speech Recognition (VSR), has been a focus of interest during the last decades [5].

Nowadays, remarkable results have been achieved in the field of VSR [4, 6]. Both the design of end-to-end architectures and the availability of large-scale databases have been fundamental pillars of recent advances in the field [5]. However, the VSR task remains an open research problem where, by dispensing with the auditory sense, challenges such as visual ambiguities and the complexity of modeling silence must be faced. In fact, Duchnowski et al. [7] maintained that only 30% of speech information is visible. Furthermore, it has been proven that each person produces speech in a unique way [8]. This fact supports the idea that visual speech features are considered as highly sensitive to the identity of the speaker [9], which poses an additional challenge when estimating speaker-independent VSR systems.

Nonetheless, these challenges can be alleviated when the VSR task is approached from a speaker-dependent perspective [10, 11]. As detailed in Section 2, there is a wide range of works which have studied the speaker adaptation of end-to-end systems in the field of Acoustic Speech Recognition (ASR) [12, 13, 14, 15]. On the contrary, few works in this regard have been addressed in VSR [16, 17]. Albeit this speaker-dependent approach means facing a less demanding task, it should not be forgotten that speaker-adapted VSR systems could be helpful, in a non-invasive and inconspicuous way, for people who suffer from communication difficulties [18, 19].

On the other hand, addressing VSR for languages other than English is recently receiving an increasing interest [20, 21, 4]. However, in these cases, the lack of audio-visual resources that it entails [22, 5] must be taken into account.

Contributions: in this paper, a study related to VSR for continuous Spanish is presented. The LIP-RTVE database [22] was used throughout all our experiments. All the defined VSR systems were based on a hybrid CTC/Attention architecture [4], which was pre-trained with hundreds of hours of data. Then, we studied how the estimation of specialized end-to-end systems for a specific person affects the quality of speech recognition. Hence, different adaptation strategies based on the fine-tuning technique were proposed. Our findings showed that significant improvements could be obtained when addressing the speaker adaptation through a two-step fine-tuning process, where the VSR system is first adapted to the task domain. Furthermore, even when only a limited amount of data was available, results comparable to the current state of the art were reached.

2. Related Work

This section offers a brief overview of the VSR task in the literature, as well as of the different works that have addressed the speaker adaption problem for speech recognizers based on end-to-end architectures. Finally, research on VSR for the Spanish language to date is considered.

Visual Speech Recognition: influenced by the evolution of systems focused on acoustic speech recognition, different approaches were considered in the field [5]. Nowadays, the current state of the art in VSR [4, 6] has shown remarkable advances, achieving around a 70% of word recognition rate on the challenging LRS3-TED database [23]. This has been possible not only to the availability of large-scale databases, but also to the design of appropriate architectures and the definition of adequate optimisation methods [4, 5].

Speaker Adaptation of End-to-End Architectures: although speaker adaptation has been widely studied with tradi-

10.21437/IberSPEECH.2022-9

tional paradigms [24, 25], in this section we only consider, due to the nature of our VSR system, those works that addressed the problem with end-to-end architectures. A simple retraining-based adaptation was adopted in [12] to fine-tune an Attention-based system. Besides, influenced by research done on conventional ASR systems [25], a hybrid CTC/Attention model was adapted by the incorporation of speaker identity vectors [13]. On the other hand, some works [14, 15] proposed the use of more sophisticated techniques such as the Kullback-Leibler divergence and Linear Hidden Networks to adapt a pre-trained speaker-independent system.

However, most of these studies were conducted in the ASR domain. Albeit this research describes approaches that could be adopted to any speech modality, it is noteworthy that few works have focused specifically on speaker adaptation for VSR systems. Kandala et al. [16] defined an architecture based on the Connectionist Temporal Classification (CTC) paradigm [26] where, once visual speech features were computed, a speaker-specific identity vector was integrated as an additional input to the decoder. Moreover, Fernandez-Lopez et al. [17] approached the problem indirectly, since it was studied how to adapt the visual front-end of an audio-visual recognition system. Thus, the authors proposed an unsupervised method that allowed an audiovisual system to be adapted when only the visual channel was available. Nevertheless, unlike our research, these works did not addressed natural continuous VSR because their experiments were evaluated on databases that were recorded in controlled settings.

Spanish Visual Speech Recognition: although they still suffer from a lack of audiovisual resources, other languages besides English are beginning to be considered in the field [27, 4]. This is the case of the Spanish language which, despite the fact that an evaluation benchmark has not yet been specified, has been the object of study on multiple occasions. Fernandez-Lopez [28] explored diverse approaches over the VLRF corpus [20], achieving around 30% accuracy at word level in the best setting of their experiments. Besides, Ma et al. [4] designed a hybrid CTC/Attention architecture that, after being pre-trained with large-scale English corpora, was fine-tuned with the Spanish CMU-MOSEAS database [27], achieving a word error rate of approximately 45%.

Regarding our previous work, we carried out several experiments where different visual speech representations were studied [21]. Subsequently, we compiled the challenging LIP-RTVE database [22], an audiovisual corpus primarily conceived to deal with the Spanish VSR task and whose details are described in Section 4. Additionally, baseline results were reported in [22], using a traditional paradigm based on Hidden Markov Models [24]. More specifically, we reached around 80% of word error rate for the speaker-dependent partition, but we were not able to obtain acceptable results (about 95% error rate) for the speaker-independent scenario. Finally, parallel to this study, we have improved the aforementioned performances in the order of roughly 40% for both scenarios, using the pre-trained CTC/Attention architecture publicly released by [4].

3. Proposed Study

The purpose of our research is to study the feasibility of developing speaker-adapted systems for the VSR task. In our case, we are going to consider the speaker dependent scenario defined for the LIP-RTVE database. Thus, we analyze how the estimation of specialized end-to-end systems for a specific person could affect the quality of speech recognition in this scenario. Hence, three different training strategies were proposed:

- **Multi-Speaker Training (MST):** the VSR system is re-estimated using the whole speaker-dependent training data; it can be seen as a task adaptation.
- **Speaker-Adapted Training (SAT):** in this strategy only data corresponding to a specific speaker is considered when fine tuning the VSR system.
- **Two-Step Speaker-Adapted Training (TS-SAT):** as its name suggests, this method consists of two fine-tuning steps. First, following the MST strategy, training data of the whole set of speakers is used to re-train the VSR system and achieve task adaptation. Afterwards, the system is fine-tuned to a specific speaker using her/his corresponding data.

In this way, by using a speaker-dependent database, we are able to study how a VSR system can generalize common patterns from different speakers or, on the contrary, evaluate to what extent it is capable of adapting to a specific speaker.

Furthermore, in order to estimate all these VSR systems, we used a simple retraining-based method, also known as fine-tuning, whose details are specified in Subsection 6.5. This decision was mainly influenced by our GPU memory constraints described in Subsection 6.8.

4. The LIP-RTVE Database

One of the main reasons why we have chosen the LIP-RTVE database[1] is because it offers, for the Spanish language, a suitable support to estimate VSR systems against realistic scenarios. It is a challenging database compiled from TV newscast programmes that were recorded at 25 fps with resolution of 480×270 pixels. No type of restriction was considered in data collection, being able to find the so-called spontaneous speech phenomena, as well as different lighting conditions or head movements. The corpus is composed of 323 speakers, providing 13 hours of data with a vocabulary size of 9308 words.

5. Model Architecture

The model architecture employed in our research is based on the work developed in [4]. The following modules are distinguished:

- **Visual Front-end:** it consists of a 2D ResNet-18 [29] whose first layer, in order to deal with data temporal relationships, has been replaced by a 3D convolutional layer.
- **Conformer Encoder:** a 12-layer Conformer [30] block is defined to capture global and local speech interactions from the previous latent visual representation.
- **Hybrid CTC/Attention Decoder:** it is composed of a 6-layer Transformer [31] block and a fully connected layer as the CTC-based decoding branch [26].
- **Language Model:** a character-level language model (LM), composed of 6 Transformer [31] layers, is integrated during the decoding process.

The entire model is estimated according to a loss function that combines both the CTC and the Attention paradigm, an approach that has led to advances in speech processing [32, 33].

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

6. Experimental Settings

6.1. Data Sets

The LIP-RTVE database defines a partition both for a speaker-dependent and speaker-independent scenario [22], each with its respective training, development and test sets. Due to the nature of the proposed study, we decided to use the speaker-dependent partition throughout our experiments. Nonetheless, in order to estimate our speaker-adapted VSR systems and properly interpret the obtained results, the 20 most talkative speakers were selected. Thus, albeit the MST-based VSR system was trained using the entire speaker-dependent partition (roughly 9 hours of data), each SAT-based model used only the data of its corresponding speaker (about 15 minutes for training and 3 minutes for development, in average terms per speaker).

6.2. Region of Interest Extraction

By using state-of-the-art open source resources[2,3] [34, 35], gray-scale bounding boxes centered on the speaker's mouth were cropped as images of 96×96 pixels. The regions not only cover the mouth, but the complete jaw and the cheeks of the speaker, which has shown benefits when addressing VSR [36].

6.3. Data Augmentation

The data augmentation process was defined according to the guidelines described in [4]. First, once the ROIs were normalized with respect to the overall mean and variance of the training set, a random cropping of 88×88 pixels was applied. Then, additional techniques such as horizontal flipping and time masking were considered.

6.4. Pre-Training

The parameters publicly released by Ma et al. [4] for the Spanish CMU-MOSEAS database [27] were used to pre-train both the VSR system and the LM described in Section 5. Concretely, the VSR system was pre-trained on over a thousand hours of data collected from multiple large-scale English corpora. Subsequently, by using the Spanish part of the CMU-MOSEAS and the Multilingual-TEDx database [37], the system was fine-tuned to the Spanish language. Regarding the LM, it was estimated from text collected from different databases until gathering more than 200 million characters.

6.5. Training & Decoding Setup

Our learning method consists of a simple retraining-based adaptation. Thus, the pre-trained VSR system was fine-tuned during 5 epochs, using the AdamW optimiser[4] and a linear One Cycle Learning Rate Scheduler[5]. The learning rate was set to 5×10^{-4} and the batch size, due to the memory constraints described in Subsection 6.8, had to consider only one sample. The rest of the hyperparameters related with the training process kept the default configuration, as detailed in [4].

In the same way, most of the decoding hyperparameters kept their default settings [4]. Nonetheless, due to our memory constraints, the beam size value was reduced to 10.

[2] https://github.com/hhj1897/face_alignment

[3] https://github.com/hhj1897/face_detection

[4] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html

[5] https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.OneCycleLR.html

6.6. Methodology

Before reporting and discussing our results, different aspects on how we carried out our experiments must be clarified:

- Both the MST- and the SAT-based systems were initially pre-trained with the parameters released by Ma et al. [4], as described in Subsection 6.4.
- The MST-based system was fine-tuned using the entire speaker-dependent partition of the LIP-RTVE database.
- A SAT-based system was independently estimated for each speaker considered in our study, i.e., 20 SAT-based systems were defined.
- Regarding the TS-SAT strategy, the previously estimated MST-based system was used as a starting point, which could be considered as an adaptation of the model to the task. Afterwards, we followed the same fine-tuning scheme described with the SAT strategy to obtain a TS-SAT-based system for each speaker.
- Experiments were conducted using either the training or the development set for adapting, as reflected in Table 1. However, it should be noted that TS-SAT-based systems were always based on the MST-based system estimated with the training set, while in the second step training or validation were used to fine-tune depending on the experiment.
- All these VSR systems, as Figure 1 shows, were evaluated on the test set corresponding to each of the speakers selected in our study. The MST-based system was the same regardless the evaluated speaker. Conversely, for the rest of strategies, the corresponding speaker-adapted system was used in each case.
- The LM used in all the tests was the one described in Subsection 6.4.

6.7. Evaluation Metric

All the results reported in our experiments were evaluated by the well-known Word Error Rate (WER) with 95% confidence intervals obtained by the bootstrap method as described in [38].

6.8. Implementation Details

The VSR system was implemented using the PyTorch[6] backend of the open-source ESPNet toolkit [39]. Experiments were conducted on a GeForce RTX 2080 GPU with 8GB memory.

7. Results and Discussion

Our first experiments were focused on the training setup. Therefore, several learning rates were explored until the optimal value, specified in Subsection 6.5, were reached. This optimum turned out to be the same for all the proposed adaptation strategies. Moreover, we studied dispensing with the scheduler, concluding that its absence not only slowed down the learning process, but also worsened the VSR system performance.

Once determined the best setting, all the VSR systems described in Subsection 6.6 were estimated. In this way, as reflected in Table 1, we could compare the proposed adaptation strategies in general terms. As mentioned above in Subsection 6.6, we explored the use of different data sets when applying our fine-tuning strategies. Thus, we were able to study how these strategies behave based on the amount of data available, since

[6] https://pytorch.org/

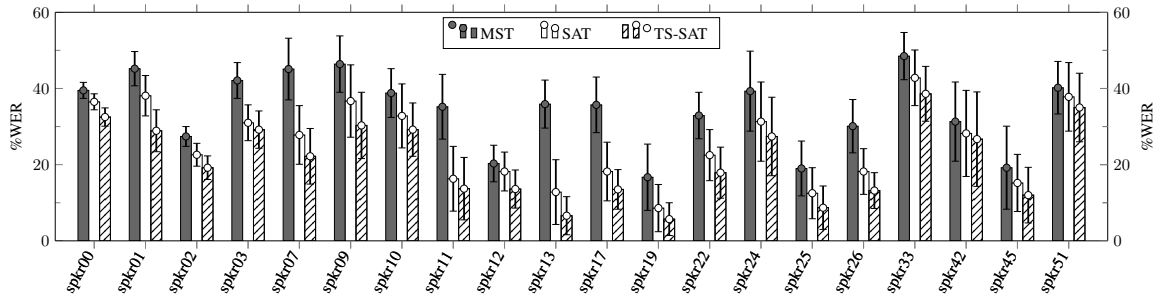

Figure 1: *Comparison of the proposed adaptation strategies. System performance (WER) with 95% confidence intervals is reported for each speaker considered in the study. Only experiments that used the training data set to estimate the VSR systems are considered.*

dealing with the development set means, in average terms, using an amount of data 4.5 times less than when using the training set [22]. By taking into account all these aspects and the results reported in Table 1, we could infer the following conclusions:

- First, regardless the fine-tuning data set used, we can observe how the MST method is significantly outperformed by the rest of the proposed strategies, a fact that supports the effectiveness of our speaker adaptation approaches.
- Nonetheless, when the training set was used, the MST-based system provided a considerable quality of speech recognition. This outcome could mean that the architecture employed in our experiments was able to generalise common patterns across speakers when addressing VSR.
- Regarding the amount of data used during the fine-tuning process, results reflect a drastic deterioration of system performance when the development set was used. However, this deterioration is noticeably lower when the TS-SAT strategy is applied, showing that this approach could be more robust against those situations in which a speaker presents data scarcity.
- Thus, the TS-SAT strategy stands as the best option when speaker adaptation is addressed. This fact supports the idea that a two-step fine-tuning process, where the model is first adapted to the general task, could benefit the final adaptation of the VSR system to a specific speaker.
- Finally, we consider that our results are comparable to the current state of the art in the field [4]. Moreover, our findings suggest that the fine-tuning method employed in our experiments is capable of adapting VSR end-to-end architectures in a small number of epochs, even when only a limited amount of data is available.

For reference, it should be noted that about 60% WER was obtained for the speaker-independent scenario.

Table 1: *System performance (WER) in average terms for each proposed adaptation strategy, depending on the data set used to fine-tune the VSR system. DEV and TRAIN refer to the development and training data set, respectively.*

Strategy	**Fine-Tuning Data Set**	
	DEV	**TRAIN**
MST	59.6±1.3	36.4±1.3
SAT	52.2±1.4	29.1±1.5
TS-SAT	32.8±1.3	24.9±1.4

On the other hand, considering only those experiments where the training set was used, we evaluated each strategy for each of the speakers selected in our study, as Figure 1 shows. From these results we could infer similar conclusions to the aforementioned ones. Nonetheless, it is noteworthy that, regardless the type of strategy applied, the VSR system provides remarkably different recognition rates depending on the speaker evaluated. Hence, a study was carried out with the aim of finding out the reasons that could explain this behaviour. For each of these speakers, several statistics were computed, such as the number of words per utterance, the perplexity of the LM in the test samples, or the number of training seconds. Thus, it was analyzed how, as the word error rate increased, each of the statistics varied. Nonetheless, we were not able to identify any trends or patterns from the data. Therefore, we could say that these experiments suggested that the reason why VSR systems behaved in this way could be related to aspects that are difficult to model, such as better vocalizations or certain oral physiognomies that reflect more adequately speech articulations.

8. Conclusions and Future Work

In our research, the continuous VSR for the Spanish language has been addressed from a speaker-adapted perspective. The challenging LIP-RTVE database was used throughout all our experiments. The VSR systems defined in our work were based on a hybrid CTC/Attention architecture [4]. Thus, it was studied how the estimation of specialized end-to-end systems for a specific speaker could affect the quality of speech recognition. Hence, different speaker adaptation strategies based on the fine-tuning technique were proposed. Our findings showed that a two-step fine-tuning process, where the VSR system is first adapted to the task domain, provided significant improvements when the speaker adaptation was addressed. Furthermore, results comparable to the current state of the art [4] were reached even when only a limited amount of data was available.

Regarding our future work, we consider exploring more sophisticated approaches as those described in [14, 15, 13, 12] when more powerful GPUs are available. Additionally, we propose to study how using auxiliary tasks, as proposed in [4], could benefit the quality of speech recognition. The use of distillation-based learning methods [40, 41] to estimate a simpler model with similar performance is another line to follow. Besides, we consider the integration of a previous speaker identification module [42]. Thus, unlike our experiments where a perfect speaker classifier was assumed, we will be able to develop a system aimed at a realistic application.

9. Acknowledgements

This work was partially supported by Generalitat Valenciana under project DeepPattern (PROMETEO/219/121) and by Grant PID2021-124719OB-I00 funded by MCIN/AEI/10.13039/501100011033/ ERDF, EU.

10. References

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

[2] B. Juang, "Speech recognition in adverse environments," *Computer Speech & Language*, vol. 5, no. 3, pp. 275–294, 1991.

[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[4] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *arXiv preprint arXiv:2202.13084*, 2022.

[5] 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.

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

[7] P. Duchnowski, D. S. Lum, J. C. Krause, M. G. Sexton, M. S. Bratakos, and L. D. Braida, "Development of speechreading supplements based on automatic speech recognition," *IEEE trans. on biomedical engineering*, vol. 47, no. 4, pp. 487–496, 2000.

[8] K.-Y. Leung, M.-W. Mak, and S.-Y. Kung, "Articulatory feature-based conditional pronunciation modeling for speaker verification," in *Proc. Interspeech*, 2004, pp. 2597–2600.

[9] S. J. Cox, R. W. Harvey, Y. Lan, J. L. Newman, and B.-J. Theobald, "The challenge of multispeaker lip-reading." in *AVSP*, 2008, pp. 179–184.

[10] K. Thangthai and R. Harvey, "Improving computer lipreading via dnn sequence discriminative training techniques," in *Proc. Interspeech*, 2017, pp. 3657–3661.

[11] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *ArXiv*, vol. abs/1611.01599, 2016.

[12] T. Ochiai, S. Watanabe, S. Katagiri, T. Hori, and J. Hershey, "Speaker adaptation for multichannel end-to-end speech recognition," in *ICASSP*, 2018, pp. 6707–6711.

[13] M. Delcroix, S. Watanabe, A. Ogawa, S. Karita, and T. Nakatani, "Auxiliary Feature Based Adaptation of End-to-end ASR Systems," in *Proc. Interspeech*, 2018, pp. 2444–2448.

[14] F. Weninger, J. Andrés-Ferrer, X. Li, and P. Zhan, "Listen, Attend, Spell and Adapt: Speaker Adapted Sequence-to-Sequence ASR," in *Proc. Interspeech 2019*, 2019, pp. 3805–3809.

[15] K. Li, J. Li, Y. Zhao, K. Kumar, and Y. Gong, "Speaker adaptation for end-to-end ctc models," in *IEEE SLT*, 2018, pp. 542–549.

[16] P. A. Kandala, A. Thanda, D. K. Margam, R. C. Aralikatti, T. Sharma, S. Roy, and S. M. Venkatesan, "Speaker Adaptation for Lip-Reading Using Visual Identity Vectors," in *Proc. Interspeech*, 2019, pp. 2758–2762.

[17] A. Fernandez-Lopez, A. Karaali, N. Harte, and F. M. Sukno, "Cogans for unsupervised visual speech adaptation to new speakers," in *ICASSP 2020*, 2020, pp. 6294–6298.

[18] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[19] K. Matsui, K. Fukuyama, Y. Nakatoh, and Y. O. Kato, "Speech enhancement system using lip-reading," in *IEEE 2nd IICAIET*, 2020, pp. 1–5.

[20] 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 *12th FG*. IEEE, 2017, pp. 208–215.

[21] D. Gimeno-Gómez and C.-D. Martínez-Hinarejos, "Analysis of Visual Features for Continuous Lipreading in Spanish," in *Proc. IberSPEECH*, 2021, pp. 220–224.

[22] ——, "LIP-RTVE: An Audiovisual Database for Continuous Spanish in the Wild," in *Proceedings of LREC*. European Language Resources Association, June 2022, pp. 2750–2758.

[23] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[24] M. Gales and S. Young, *The application of hidden Markov models in speech recognition*. Now Publishers Inc, 2008.

[25] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *IEEE ASRU*, 2013, pp. 55–59.

[26] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd ICML*. ACM, 2006, p. 369–376.

[27] A. B. 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 EMNLP*, 2020, pp. 1801–1812.

[28] A. Fernandez-Lopez and F. M. Sukno, "End-to-end lip-reading without large-scale data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2076–2090, 2022.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[30] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[33] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *IEEE SLT*, 2018, pp. 513–520.

[34] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *IEEE/CVF CVPR*, 2020, pp. 5202–5211.

[35] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *IEEE ICCV*, 2017, pp. 1021–1030.

[36] Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition," in *15th IEEE FG*, 2020, pp. 356–363.

[37] 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, pp. 3655–3659.

[38] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in *ICASSP*, vol. 1. IEEE, 2004, pp. 409–412.

[39] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[40] J. W. Yoon, H. Lee, H. Y. Kim, W. I. Cho, and N. S. Kim, "Tutornet: Towards flexible knowledge distillation for end-to-end speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1626–1638, 2021.

[41] T. Afouras, J. S. Chung, and A. Zisserman, "ASR is all you need: Cross-modal distillation for lip reading," in *ICASSP*, 2020, pp. 2143–2147.

[42] M. Uzun-Per and M. Gökmen, "Face recognition with patch-based local walsh transform," *Signal Processing: Image Communication*, vol. 61, pp. 85–96, 2018.