EURASIP Journal on Audio,
Speech, and Music Processing

EMPIRICAL RESEARCH **Open Access**

Continuous lipreading based on acoustic temporal alignments

David Gimeno-Gómez[1*†] and Carlos-D. Martínez-Hinarejos[1†]

Abstract

Visual speech recognition (VSR) is a challenging task that has received increasing interest during the last few decades. Current state of the art employs powerful end-to-end architectures based on deep learning which depend on large amounts of data and high computational resources for their estimation. We address the task of VSR for data scarcity scenarios with limited computational resources by using traditional approaches based on hidden Markov models. We present a novel learning strategy that employs information obtained from previous acoustic temporal alignments to improve the visual system performance. Furthermore, we studied multiple visual speech representations and how image resolution or frame rate affect its performance. All these experiments were conducted on the limited data VLRF corpus, a database which offers an audio-visual support to address continuous speech recognition in Spanish. The results show that our approach significantly outperforms the best results achieved on the task to date.

Keywords Visual speech recognition, Limited computation, Data scarcity, Speech processing, Computer vision

1 Introduction

Speech is considered a process where multiple senses are involved, including high-level knowledge such as grammar, semantics, and pragmatics [1]. Our brain is responsible for integrating all this information to improve our ability to understand the message we are perceiving. Furthermore, different studies [2, 3] showed the relevance of visual cues during our speech perception process. For instance, McGurk and MacDonald [3] demonstrated that if the mouth expression does not match the emitted sound, the listener is confused, perceiving a sound different from what it was.

In its origins, automatic speech recognition (ASR) was focused solely on acoustic cues [4, 5]. Nowadays, auditory-based ASR systems are capable of understanding spoken language with great quality [6–8]. However, the performance of these systems considerably deteriorates in noisy environments, where the acoustic signal could be damaged or corrupted [9–11]. Therefore, inspired by our multi-sensory process, different approaches [7, 10, 12, 13] were designed from an audio-visual perspective to improve the robustness of ASR in such adverse scenarios. These approaches addressed the so-called audio-visual speech recognition (AVSR) task, whose architectures are considered the current state of the art in the field [7, 10, 12, 14]. Other lines of research are focused on speech enhancement methods [15–18] without the need to rely on complementary visual cues. Nonetheless, there has been an increasing interest in visual speech recognition (VSR) during the last few decades [19]. Specifically, this task, also known as automatic lipreading, aims to interpret speech based exclusively on lip movements. Hence, different challenges must be considered by dispensing with the auditory sense, such as visual ambiguities or the complex modelling of silence [20, 21].

[†] David Gimeno-Gómez and Carlos-D. Martínez-Hinarejos contributed equally to this work.

*Correspondence:
David Gimeno-Gómez
dagigo1@dsic.upv.es
[1] Pattern Recognition and Human Language Technologies Research Center, Universitat Politècnica de València, Camino de Vera, s/n, València 46022, Comunitat Valenciana, Spain

Our research focuses on VSR since, although noticeable advances have been achieved in the field [12, 14, 22], it is still considered an open research problem. Moreover, recognizing speech without the need for acoustic stream data offers a wide range of applications, from silent speech passwords [23] or visual keyword spotting [24] to the development of silent speech interfaces that would be able to improve the lives of people who suffer from communication difficulties [25, 26].

A remarkable aspect is that these recent advances in VSR [12, 14, 22] rely on powerful end-to-end architectures based on deep learning techniques. Such approaches are not only highly dependent on large amounts of data for their estimation, but they also require extensive computational resources. However, there are situations where it is not possible to satisfy these demands. One example would be if our purpose consisted of developing a VSR mobile application for a low-resource language. These were the main reasons why we decided to construct our VSR system based on those traditional approaches originally explored in the field of ASR [4, 5, 27]. Specifically, we studied architectures based on hidden Markov models (HMMs), either combined with Gaussian mixture models (GMM-HMM) [4] or deep neural networks (DNN-HMM) [27], as well as the use of techniques such as sequence discriminative training (SDT) [28].

Unlike the recent end-to-end architectures [29], these traditional approaches require the estimation of several independent modules, as well as the use of pre-processed speech features, tree-based clustering, or different knowledge-based techniques [30]. Thus, in addition to estimating the speech recognizer module and the language model, it is necessary to define a lexical model, where words are associated with the basic units of speech that compose them. Furthermore, as reflected in Sect. 3.4, these HMM-based systems are based on elaborate training schedules composed of multiple stages, each relying on the temporal alignments previously provided by a preliminary stage. An important aspect regarding these alignments is the HMM's topology, since depending on the number of states and transitions with which it was defined, we will condition how speech features are related to basic speech units.

Therefore, defining a basic speech unit is necessary when dealing with these traditional paradigms. In the visual domain, this concept is associated with the so-called viseme [31]. Unfortunately, there is no direct or one-to-one correspondence between the audio-based phonemes and the visemes, which causes the ambiguities previously mentioned [20]. In the VSR literature, there is an open discussion about the use of phonemes or visemes. Many authors have extensively studied both viseme- [20, 32, 33] and phoneme-based approaches [34–36]. In our work, we considered phonemes as our basic speech units.

Another important aspect is the temporal alignment of our data. HMM-based systems are estimated using a sequence of different stages, where each stage relies on the temporal alignments provided by the previous stage. However, it should be noted that visual data usually presents a lower sample rate than audio data. Therefore, the optimal HMM's topology can be different from that used in auditory-based ASR, and it must be considered as a critical element in our experiments.

Other details related to lipreading complexity are the intra-personal variability among speakers, the different light conditions, or technical characteristics, such as frame rate or image resolution [37, 38]. For all these reasons, a suitable feature extraction becomes a fundamental pillar. Early works focused their research on this regard using conventional data transform techniques [39, 40] or the so-called autoencoders [41], without reaching any consensus on what was the best visual speech representation [19]. Nowadays, these conventional approaches have been eclipsed by the rise of self-supervised encoders based on powerful attention mechanisms [14, 42] that present high data and computational requirements. Consequently, since we are focusing on data and computationally limited scenarios, we explored three different visual speech features based on more conventional approaches that do not have high data and computational demand.

Albeit visual cues (through the movement of lips, teeth, and tongue) can provide valuable information, a study carried out by [43] supported that only around 30% of speech information is visible. Hence, as described in Sect. 2, the combination of acoustic and visual cues has been extensively studied not only to address AVSR [7, 13], but also to enhance one modality using the complement knowledge from the other [14, 42]. Influenced by all these works, this paper presents, to the best of our knowledge, a novel learning strategy (ViTAAl, see Sect. 3.3) that employs information obtained from previous auditory-based temporal alignments to improve the performance of a HMM-based VSR system, as Fig. 5 reflects.

Finally, our experiments were conducted using the limited-data VLRF corpus, a speaker-dependent database that offers, although in controlled recording settings, an audio-visual support to address continuous ASR in Spanish. Details on the database can be found in Sect. 3.1.

1.1 Contributions

All these were the main reasons that motivated our research, where our key contributions are (i) addressing the task of continuous VSR not only for data scarcity scenarios, but also when computational resources are limited; (ii) the proposal of the ViTAAl learning strategy, whose purpose is to improve the performance of traditional HMM-based VSR systems by the use of auditory-based temporal alignments; (iii) an analysis on how the effectiveness of our proposed strategy is affected by the image resolution and the frame rate; (iv) a comparative study on three different visual speech representations based on conventional approaches; and (v) results show that our approach significantly outperforms the best results achieved on the task to date, demonstrating that visual cues benefit when acoustic-based knowledge is incorporated when estimating traditional VSR systems.

2 Related work

This section presents a brief overview of various aspects related to our research, such as the use of traditional paradigms, the current state of the art in the field, how numerous approaches have integrated acoustic and visual cues with different purposes, and the study of the task regarding the Spanish language.

2.1 Traditional VSR approaches

Although there was extensive research on VSR using HMMs, most of these early works focused on simpler tasks, such as alphabet or digit recognition [19]. Regarding our interest in continuous VSR, Thangthai et al. [36] studied both GMM-HMM and DNN-HMM models based on phonemes for the single-speaker RM-3000 corpus [34]. In this case, the authors employed active appearance models [44] to extract the features related both to the shape and appearance of the mouth region. Subsequently, Thangthai and Harvey [45] evaluated their systems on the TCD-TIMIT corpus [46], using the so-called eigenlips [44, 47] as visual speech features. The results were reported for context-dependent GMM-HMM in speaker-dependent and speaker-independent scenarios, achieving around 71.2% and a 75.4% word error rate (WER), respectively. DNN-HMM models were also explored, representing an improvement of around 25% in both cases. Additionally, different data transform techniques widely used in the field of ASR [4] were also applied throughout the training phases.

2.2 Current VSR approaches

Most recent works [7, 14, 48], as well as events such as the MISP challenge [49], focused primarily on the AVSR task. However, there has been an increasing interest in VSR [19]. Shi et al. [14] introduced AV-HuBERT, a self-supervised audio-visual encoder capable of learning robust visual speech representations. This encoder was then assembled with a transformer-based decoder to build an end-to-end VSR system. Prajwal et al. [50] not only proposed an attention module aimed explicitly at extracting representative visual features, but also explored sub-word units, arguing that it might be helpful for better modelling visual ambiguities. Ma et al. [22] showed in their study that, apart from the importance of hyperparameter optimization and data augmentation, incorporating auxiliary tasks to an end-to-end architecture might lead to further advances in the field. In general terms, all these works reached performances around 25–30% WER for the English corpora LRS2-BBC [51] and LRS3-TED [52].

2.3 Audio-visual integration

The multimodal integration of acoustic and visual cues is a research field that has aroused great interest in multiple and varied tasks for a long time [53–55]. In the automatic speech recognition domain, the combination of acoustic and visual cues has also been extensively studied for decades [14, 56]. In the field of AVSR, as Potamianos et al. [13] discussed, several authors evaluated different methods regarding audio-visual speech integration, either through the direct fusion of features [14, 22, 57, 58] or by employing approaches that fuse the knowledge of two classifiers, one for each modality [1, 56, 59, 60]. Other authors have explored methods to learn shared audio-visual representations [14, 54] or visual-based speech features enhanced by information extracted from acoustic cues [42, 60, 61]. The approach presented in [62] is based on multi-task learning [63]. It proposes a shared representation DNN-HMM which, receiving acoustic and visual data as input, attempts to predict, for both modalities, the appropriate sequence of HMM's states. In order to obtain the required target labels to estimate this system, the authors previously trained an independent GMM-HMM for each modality. Other interesting works were presented in [59, 60], where an end-to-end VSR model was trained by distilling from an auditory-based ASR model in a teacher-student manner. All these works significantly influenced our proposed ViTAAl strategy.

2.4 VSR for Spanish

As previously mentioned, the current state of the art in VSR is around 25–30% WER for the widely studied English benchmark [14, 22]. However, this recognition rate is notably reduced when addressing languages other than English [22]. In the case of Spanish, our language of interest, different works have been carried out in the field of VSR [22, 64–68]. Nonetheless, it has been the work published by [22] the one which, by proposing a powerful end-to-end architecture, has reached a new state of the art for the task, reporting results of around 50% WER for different Spanish corpora.

Regarding the VLRF corpus, Fernandez-Lopez et al. [64] designed a viseme-based system that achieved around 80% WER. As visual speech features, they considered combining the descriptors obtained by the discrete cosine transform [69] and scale-invariant feature transform [70] methods. However, their system does not correspond with the conventional paradigm in ASR. In further research, Fernandez-Lopez and Sukno [65] proposed an end-to-end approach capable of achieving, in the best setting along their experiments, a 72.9% WER. It should be noted that data augmentation techniques were applied to overcome the limited data offered by the database.

3 Methodology

3.1 The VLRF corpus

One of the main reasons why we have chosen the visual lip-reading feasibility (VLRF) corpus [64] is because it offers, although in controlled recording settings, a public audio-visual support to address continuous VSR in Spanish (our language of interest) for data scarcity scenarios. Furthermore, various studies have presented baseline results using conventional HMM-based frameworks [64] and more recent end-to-end architectures [65], allowing us to conduct a thorough comparison of our proposed method and its effectiveness with different speech recognition paradigms.

This speaker-dependent corpus was designed to study the feasibility of the task, so speakers were asked to strive to be understood. The database was acquired with the participation of 24 people, each one with 25 assigned unrelated sentences. Each one of these sentences was repeated 3 times in a recording studio with controlled lighting conditions. In this studio, a camera captured video at 50 fps with a resolution of 1280 × 720 pixels, while audio was recorded at 48 kHz mono with 16-bit precision. It should be noted that, as specified in Sect. 3.5, only the first repetition of each sentence was considered when estimating an ASR system [64, 65]. Therefore, the VLRF corpus contains a vocabulary of 1374 words and a duration of around 1 h of data.

The VLRF corpus represents the ideal setting for our case study, whose main purpose is to address the VSR task for data scarcity scenarios with limited computational resources, particularly valuable aspects when developing a mobile application. An interesting example, recently proposed by Liopa[1], would be the development of a mobile application aimed to help speech-impaired people who suffer from difficulties in speaking due to tracheostomy, laryngectomy, stroke, or other types of injuries.

3.2 Feature extraction

3.2.1 Acoustic speech features

As in traditional ASR, we apply the standard representation based on Mel frequency cepstral coefficients (MFCC) along with first- and second-order dynamic differential parameters (Δ+$\Delta\Delta$) [30]. Nevertheless, visual data usually presents a lower sample rate than audio data [13]. This fact would cause conflicts when audio and video are combined. For this reason, in the field of AVSR, several authors decided to up-sample visual data to adjust this modality with the frame rate of acoustic features [36, 45, 71].

However, we have considered it a better option to adapt the acoustic stream to the sample rate of visual data. Therefore, we extract the 39-dimensional MFCC+Δ+$\Delta\Delta$ [30] but with a frame shift of 20 ms that fits the 50 fps video capturing for VLRF.

3.2.2 Visual speech features

As described in Sect. 1, the development of our VSR system aims at scenarios where data and computational resources are limited. For this reason, although powerful attention-based encoders have led to advances in the field [14, 22], we considered more conventional approaches in our experiments. However, as there is no consensus on what is the best option among these conventional approaches to represent the nature of visual speech [19], we studied three different types of visual features as we describe below.

In all the studied approaches, the use of the OpenCV library [72] and the Dlib toolkit [73] allowed us to identify 68 facial landmarks [74]. From some of these landmarks, we were able to extract our region of interest (ROI), delimited by the green box in Fig. 1. Subsequently, all these ROIs were converted to grey-scale images, normalized to a resolution of 128 × 64 pixels, and, in addition, a histogram equalization was applied to them.

- Geometric features: In this first approach, our main inspiration was the work carried out in [75] and [76]. We defined, based on landmarks locations, a set of 19 high-level features, such as width, height, or area of the speaker's mouth, as reflected in the third image of Fig. 1. Additionally, in order to compute features as stable as possible against any movement of the speaker, we normalized each measure regarding the dimensions of a less variable region, highlighted by a larger blue rectangle on the second image of Fig. 1.
- Eigenlips: Albeit its origin resides in the studies focused on facial recognition [77], this concept has been widely explored in VSR [44, 47]. In our work, principal component analysis (PCA) [78] was applied over a data set made up of 25 random frames for

[1] https://liopa.ai/sravi-app/

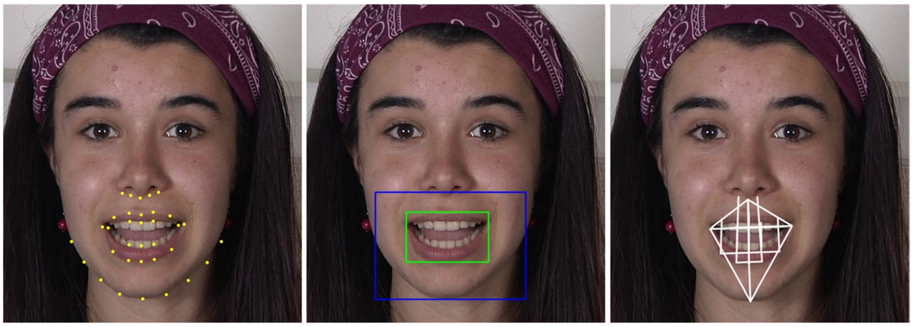

Fig. 1 Aspects regarding geometric features

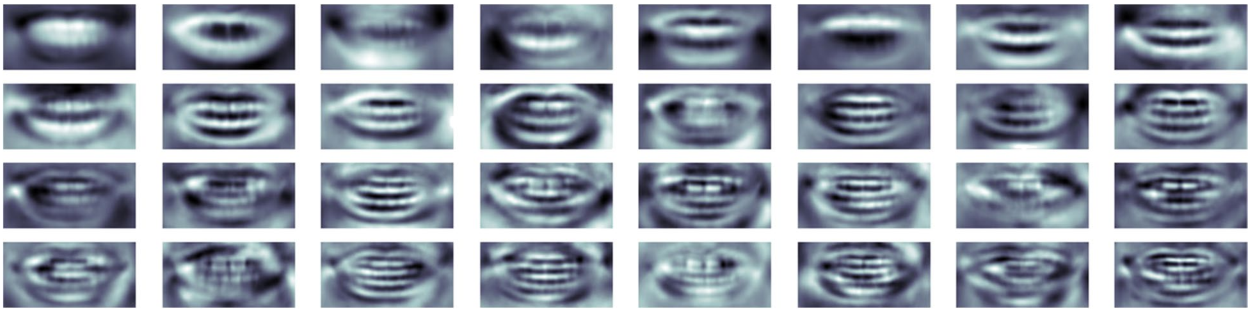

Fig. 2 The eigenlips obtained after applying PCA

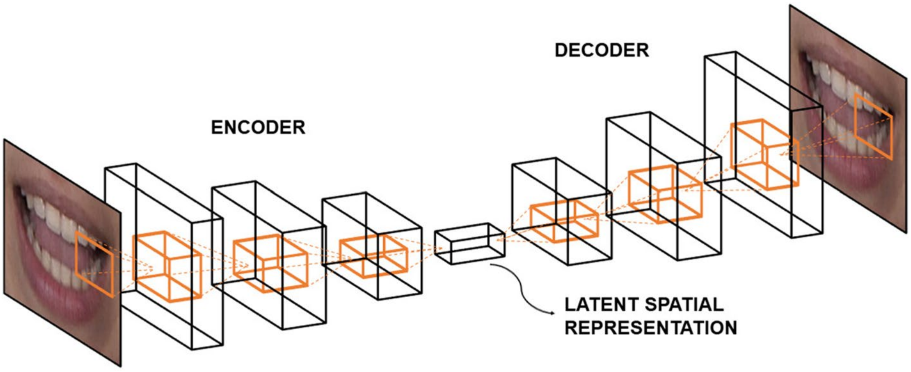

Fig. 3 Scheme of the convolutional autoencoder

each training sample, obtaining 32 eigenlips. As Fig. 2 shows, each component highlights different appearance aspects, such as lip contours or zones where we can find teeth or tongue. These are aspects that we could not reach when we used pure geometric features.

- Deep features: The last approach, as many authors have studied [41, 79, 80], is based on convolutional autoencoders (CA) [81]. As Fig. 3 depicts, this type of neural networks aims to reconstruct the input image from an abstract and compact representation. Then, once this model has been trained, we dispense with the decoder section because it is the encoder the component we need to extract our visual speech features. The encoder architecture is entirely based on that presented by [79]. Nonetheless, it was necessary to adjust our ROI's resolution by means of stacking 2 additional convolutional layers with a stride of value 2. We estimate our CA using a data set composed of 200 random frames for each training sample, employing a 32-dimensional latent spatial representation. In this way, we

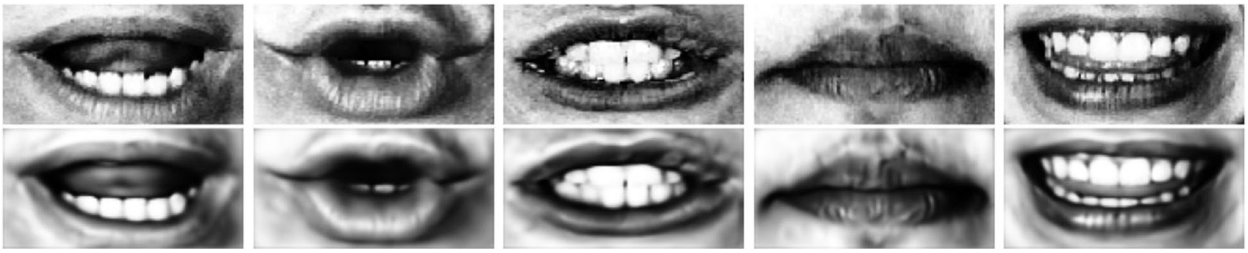

Fig. 4 Reconstruction examples obtained by the defined convolutional autoencoder. Top row: original images; bottom row: reconstructions

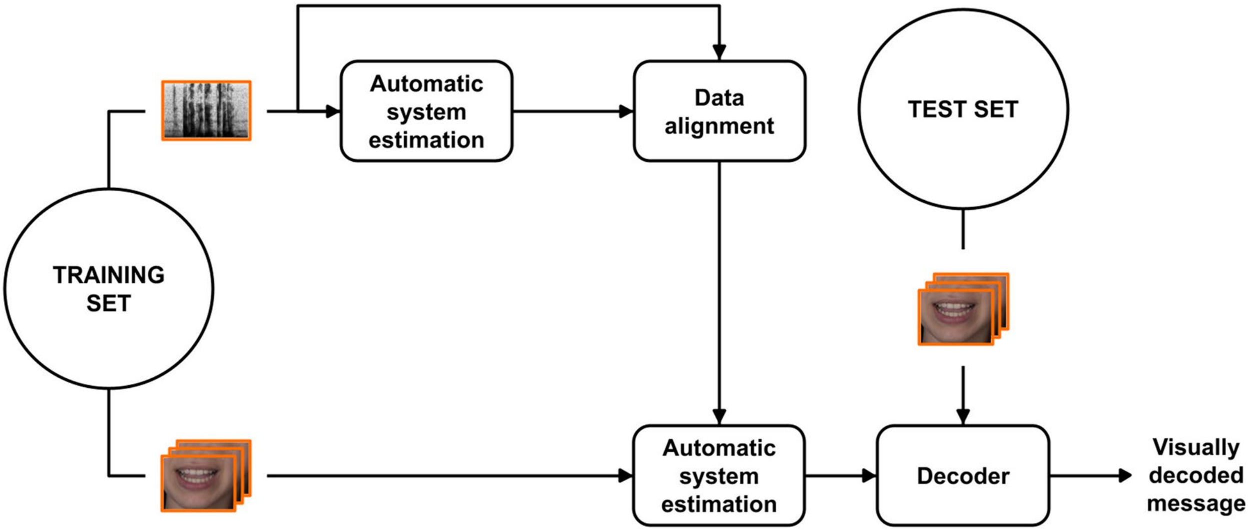

Fig. 5 Scheme of the ViTAAl strategy

were able to obtain high-quality reconstruction results, as it is shown in Fig. 4.

3.3 Visual training based on audio alignments (ViTAAl)

As described in Sect. 2, the audio-visual speech integration process has been extensively studied for decades [13, 42, 54, 59, 62]. Inspired by all these works, we designed the visual training based on audio alignments (ViTAAl) strategy. To the best of our knowledge, it constitutes a novel contribution to improving the performance of HMM-based VSR systems by using the information embedded in acoustic temporal alignments.

Once we have adequately extracted both acoustic and visual speech features, we are able to describe the proposed strategy. As Fig. 5 reflects, this process is composed of several phases, distinguishing two main blocks:

1. Initially, we build the audio-only module. Therefore, we estimate an automatic HMM-based system from scratch, either based on GMMs or DNNs, as we subsequently describe in Sect. 3.4. To do this, we use the acoustic speech features corresponding to the training set. Then, we compute the HMM state-level temporal alignments of acoustic data, which indicate approximately when each spoken phoneme begins and ends.
2. Thereafter, we define a different HMM-based system for the corresponding visual data. In this case, however, we do not estimate it from scratch but by using the previously obtained acoustic temporal alignments. These alignments not only provide the model with prior knowledge that helps it converge more easily, but also improve the extracted visual speech features. It is possible thanks to the HiLDA feature transformation [82] applied during the training process and whose details are described in Sect. 3.4

Our method allows the estimation of a VSR system capable of integrating acoustic information without the need to increase the number of parameters composing the model, and which can perform inferences on a visual-only test set. Due to these alignments being defined by classifying each frame in feature sequence into its corresponding HMM state, they are modality-independent and therefore easily integrable into another HMM-based model. Experimental results reported in Sect. 4 support the effectiveness of our proposed method, demonstrating that these acoustic-based alignments can be considered a suitable foundation for building VSR systems.

3.4 Speech recognition system

The ASR systems employed in our research were implemented using the Kaldi toolkit [83], where several workflows or recipes to build different paradigms in the field of speech technologies are provided. Specifically, our final objective was to design a hybrid DNN-HMM system based on the *Wall Street Journal* (WSJ) recipe[2]. However, this process implies several stages that obtain as by-product intermediate systems that would be tested as well. Note that each intermediate system relies on the temporal alignments provided by the previous one, so the ASR model is trained through an incremental refinement process. It allowed us to investigate at which training stage our proposed ViTAAl strategy could be more effective. This model architecture was used during our experiments for both audio and video modalities, considering in each case the corresponding speech features described in Sect. 3.2.

3.4.1 *GMM-HMM system*

The first step is to estimate a conventional GMM-HMM system, since hybrid DNN-HMM models rely on the alignments provided by this preliminary system. The GMM-HMM system is built through an incremental process composed of several phases, each one in charge of different aspects:

- MONO: a context-independent GMM-HMM is estimated from scratch; the final used features were obtained by applying over the raw features the cepstral mean and variance normalization technique and the $\Delta+\Delta\Delta$ coefficients [30].
- DELTAS: in this phase, a context-dependent GMM-HMM (using the same features than in the previous step) is trained on the basis of prior temporal alignments. Once this system is estimated, temporal alignments are updated.
- LDA+MLLT: this stage aims to reduce the feature dimensionality and capture contextual information, obtaining the known as HiLDA features [82]. Two data transform techniques are applied to compute these features. First, based on the previous DELTAS alignments, linear discriminant analysis (LDA) [84] is computed over each feature vector along with its corresponding spliced context frames, reducing it to 40 feature components. Then, a maximum likelihood linear transform (MLLT) [85] is applied. Finally, the GMM-HMM and the temporal alignments are updated.
- SAT: by applying a speaker adaptive training (SAT) [86] based on the feature space maximum likelihood linear regression (fMLLR) [4] method, the last GMM-HMM system is obtained.

3.4.2 *DNN-HMM system*

The essence of these hybrid systems is to replace GMMs with DNNs, since neural networks have shown a better capability to model data that lie in a non-linear representation space [27]. Thus, a feed-forward neural network is trained to classify each data frame into the HMM's triphone state that most likely would have emitted it. Hence, the required alignments to estimate these DNNs must be obtained from a preliminary GMM-HMM system.

In our case, we built a DNN-HMM based on the Karel's setup[3] [28]. Concretely, the first step consists of an unsupervised pre-training phase based on restricted Boltzmann machines [87]. In this way, we are able to initialize DNNs which are subsequently estimated via a frame-level cross-entropy training, employing algorithms such as mini-batch stochastic gradient descent and error back-propagation [27]. The 40-dimensional fMLLR features are employed throughout the entire process, as these decorrelated features are more suitable to be treated by DNNs [27].

3.4.3 *Sequence discriminative training*

As suggested above, DNNs are considered frame-level discriminative classifiers. However, due to the nature of speech, several authors found that DNN-HMM systems benefit if fine-tuned via sequence-level criteria [28, 88, 89], where the optimization function is directly related to target word sequences. The success of these sequence discriminative training (SDT) techniques relies on the contribution of the language model, as well as the consideration of a large context window during parameter optimization [89]. Specifically, in our research, we evaluate the most commonly studied criteria in the ASR field: maximum mutual information (MMI) [90], minimum phone error (MPE) [91], and state-level minimum Bayes risk (sMBR) [92, 93].

[2] https://github.com/kaldi-asr/kaldi/tree/master/egs/wsj/s5

[3] https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/local/nnet/run_dnn.sh

3.4.4 Decoding

Finally, the decoder is defined as a weighted finite-state transducer (WFST) [94], which integrates the morphological model, phonetic context-dependencies, the lexicon, and the language model.

3.5 Experimental setup

3.5.1 Data sets

The VLRF corpus is split as it is described in [65]. Therefore, only the first repetition of each sentence is considered, dividing them into two speaker-dependent partitions. The training set presents 55 min of data in 480 samples, while the test set contains 14 min in 120 samples.

3.5.2 Lexicon model

It was created using the transcriptions from the training and test sets, providing a 1374-word vocabulary. Specifically, this model considered 23 phonemes defined according to Spanish phonetic rules [95]. In addition, default *silence* phones of Kaldi were then included.

3.5.3 Language model

The nearly 300k sentences provided by [65] were used to train a 4-gram language model using the SRLIM toolkit [96]. In this way, the perplexity reached, with 8 out of vocabulary words, a value of 463.2. This perplexity must be taken into consideration, since it presents a significantly higher value in contrast with other works [36, 45], which evaluated perplexities at the most of 35.16.

3.5.4 Implementation details

Experiments were conducted on a 6-core 2.90 GHz Intel i5-9400 CPU with 16 GB memory. On average, the entire training schedule proposed in this paper takes around 1.5 h. Specifically, the GMM-HMM estimation took around 2 min, the DNN-HMM system needed around 1.1 h, and the SDT technique required 23.5 min. The training and decoding configuration of our recognition system is mainly based on the WSJ recipe.

- GMM-HMM training setup: First of all, different HMM's topologies were explored to study how the lower sample rate we used to fit the visual data affected the performance of the audio-based system. However, despite dealing with a lower sampling rate, the standard topology in ASR (three states left-to-right with loops topology) provided the best recognition rates. Therefore, due to the fact that a similar behaviour was found in the video-only modality and since the ViTAAl strategy is based on the quality of these acoustic alignments, only experiments using this classical topology were considered. Although the default configuration was kept for audio-only GMM-HMM experiments, different aspects were explored in our video-based scenarios. Specifically, in the case of the DELTAS training step, we studied contexts from 1 to 3 frames to compute the $\Delta+\Delta\Delta$ coefficients. For the LDA+MLLT stage, we studied contexts from 1 to 10 frames when applying HiLDA. Details on the GMM-HMM best settings for audio-only, video-only, and ViTAAl scenarios are specified in Sects. 4.1, 4.2, and 4.3, respectively.
- DNN-HMM training setup: Once the best setting for GMM-HMM systems was found, we were able to focus our experiments on the DNN-HMM hybrid architecture. The pre-training phase kept its default configuration. Conversely, a wide range of parameters has been studied in the cross-entropy training phase. Concretely, we explored different depths from 1 to 6 hidden layers, studying, in turn, values from 128 to 2048 neurons for each of these layers. In addition, the sigmoid and parametric ReLU activation functions were evaluated. The learning rate was explored from 0.01 to 0.00001 in decreasing powers of 10. Associated with this parameter, on the one hand, the momentum and halving factor were evaluated with values 0.0, 0.4, and 0.8, and, on the other hand, it was studied the effect of keeping this learning rate during the first 0, 10, or 20 iterations. It was also analysed the DNN's behaviour for an input context from 0 to 6 spliced frames. Details on the DNN-HMM best settings for audio-only, video-only, and ViTAAl scenarios are specified in Sects. 4.1, 4.2, and 4.3, respectively.
- SDT training setup: SDT was applied during 15 iterations, keeping the default setting. The 3 criteria mentioned in Sect. 3.4 were evaluated in our experiments. However, only the results provided by the sMBR technique were considered, as this criterion provided, although without significant differences w.r.t the rest of the criteria, the best recognition performance in all cases.
- Inference setup: For decoding, we set a value of 13 to pruning beam and 6 to lattice beam. The speech model scale factor was set to 0.08333, while the language model covered scale factors between 1 and 20. On the other hand, based on the BABEL recipe[4], word insertion penalty values between − 5.0 and 5.0 were applied. All these decoding parameters were evaluated in each experimental prove, but only the lowest word error rate was considered.

[4] https://github.com/kaldi-asr/kaldi/blob/master/egs/babel/s5/local/score_combine.sh

Table 1 Results (%WER) for audio-only GMM-HMM recognition depending on training phase. The best performance is in bold

Training phases			
MONO	**DELTAS**	**LDA+MLLT**	**SAT**
24.8 ± 4.3	12.6 ± 3.2	10.1 ± 3.1	**9.1 ± 3.0**

- Evaluation metric: All the results presented in this paper were evaluated by the well-known word error rate (WER) with 95% confidence intervals obtained by the bootstrap method as described in [97].

4 Experiments

This section is structured following the training scheme reflected in Sect. 3.4 when estimating HMM-based ASR systems. Hence, different aspects must be considered depending on whether it is a GMM-HMM or a DNN-HMM system and the training stage in which we find ourselves. Additionally, with the aim of proving the effectiveness of our proposed ViTAAl strategy, we first estimate the automatic systems for audio- and video-only scenarios (see Sects. 4.1 and 4.2, respectively) as our baselines. Then, an automatic VSR system is estimated based on the ViTAAl strategy (see Sect. 4.3). Finally, an overall comparison is discussed in Sect. 4.5.

It should be noted that for both the video-only and the ViTAAl systems, a comparative study was carried out on the proposed visual speech features. Moreover, due to the reasons presented in Sect. 4.3, we only considered the alignments provided by the audio-based system in the DELTAS and LDA+MLLT phases when estimating our ViTAAl-based systems. In addition, we also analysed how the effectiveness of the ViTAAl strategy was affected by the image resolution and the frame rate.

4.1 Audio-only results

As experiments reported in Sect. 4.3 support, the quality of the acoustic alignments is a fundamental aspect of our proposed strategy. Table 1 reflects how, in general terms, the acoustic-based GMM-HMM system performance improves as we progress through the training stages, achieving around 9.1% WER in the best case.

On the other hand, despite evaluating a large number of configurations, subsequent experiments regarding DNN-HMM and SDT did not improve the GMM-HMM recognition performance. This fact might suggest that, probably due to the limited available data, the lower bound has been reached for the audio-only task.

4.2 Video-only results

With the aim of proving the effectiveness of the proposed ViTAAl strategy, video-only baseline results must be obtained first. Similar to our previous audio-only experiments, we studied how the different training phases affect the GMM-HMM system performance. The best results in MONO and DELTA models were obtained, for all cases, when the delta context was 1; hence, this value was fixed for all the experiments. For the LDA+MLLT model, only the best HiLDA spliced context was considered. The results on SAT systems were disregarded, since they caused, in all cases, a slight increase in error rates.

Table 2 Results (%WER) in video-only GMM-HMM recognition for each set of features. The best training phases performances is in bold. *G* geometric features, *E* eigenlips, *D* deep features

Features	**Training phases**		
	MONO	**DELTAS**	**LDA+MLLT**
G	94.7 ± 1.3	94.7 ± 1.3	**94.2 ± 1.4**
E	**92.4 ± 2.2**	93.8 ± 2.2	92.6 ± 2.3
D	92.6 ± 2.1	93.7 ± 2.9	**91.4 ± 2.3**
G+E	**90.5 ± 2.0**	96.3 ± 2.9	91.1 ± 2.1
G+D	92.8 ± 2.1	97.3 ± 3.0	**89.1 ± 2.2**
E+D	96.7 ± 2.2	103.1 ± 3.2	**90.9 ± 2.2**
G+E+D	95.0 ± 2.6	106.3 ± 3.9	**88.4 ± 2.6**

Experiments reported in Table 2 were focused on the visual speech feature comparison along the different training steps. First, we found that DELTA systems provided poorer-quality results as feature dimensionality increased through combinations. Conversely, in general terms, LDA+MLLT models offered better results as feature dimensionality increased. This behaviour might be associated with the use of HiLDA features, since this technique is able to extract features in a discriminative way as well as to reduce their dimensionality, which eases data modelling. Furthermore, we could relate the cases in which MONO systems were the best approach to the limited amount of data of the VLRF corpus, which makes context-dependency modelling difficult for the rest of the models.

Regarding our visual feature analysis, we only consider the LDA+MLLT results. When comparing the studied features in isolation, we conclude that deep features provided the best recognition rate. In fact, we can observe how the rest of the features improved their performance considerably when they were separately combined with the deep representation. Nonetheless, the best performance was obtained when all features were combined, achieving an error rate of around 88.4% WER.

Once our best video-only GMM-HMM is found, we were able to estimate the corresponding DNN-HMM system. The experiments allowed to conclude that the best system was composed of 1 hidden layer with 256

Table 3 Results (%WER) in GMM-HMM ViTAAL recognition for each set of features depending on the training stage where the audio-based alignments were obtained. The best performance for each approach is in bold. *G* geometric features, *E* eigenlips, *D* deep features

Features	Audio alignments from	
	DELTAS	LDA+MLLT
G	95.8 ± 1.3	85.9 ± 2.8
E	**82.3 ± 2.5**	73.6 ± 3.7
D	82.4 ± 2.6	77.0 ± 3.5
G+E	85.6 ± 2.5	72.9 ± 3.9
G+D	85.9 ± 2.6	74.1 ± 3.7
E+D	91.1 ± 3.0	71.2 ± 3.8
G+E+D	91.1 ± 3.2	**68.7 ± 4.0**

neurons and sigmoid activation, reaching around 84.2% WER. The 40-dimensional fMLLR features spliced with 9 context frames were used as input. The learning rate was set to 0.001 and kept during the first 10 iterations; the halving factor and momentum were set to 0.4 and 0.0, respectively. Regarding SDT experiments, our best video-only DNN-HMM system achieved an error rate of 83.4% after applying the sMBR criterion, a 5% improvement over the GMM-HMM paradigm.

4.3 ViTAAl results

As described in Sect. 3.3, the ViTAAl strategy uses the temporal alignments provided by the acoustic module. Therefore, if we decide, for example, to use the DELTAS audio alignments, we must estimate a visual system that can take advantage of this knowledge, that is, a visual DELTAS model. However, MONO systems only support flat estimations, and SAT systems need a specific visual model to transform the visual data properly. For this reason, as Table 3 reflects, we could only evaluate our strategy with DELTAS and LDA+MLLT audio alignments. As in the video-only experiments, a one-frame context for DELTA systems was the best setting. Regarding the LDA+MLLT approach, only the best HiLDA spliced context was considered.

According to the results reported in Table 3, visual LDA+MLLT GMM-HMM systems based on the ViTAAl learning strategy outperform, in all cases, ViTAAl-DELTAS systems. This behaviour is coherent with the audio-only experiments where, as Table 1 confirms, better quality alignments are provided by the LDA+MLLT system. On the other hand, if we compare these ViTAAl results with those presented in Table 2, we observe that, in average terms, our proposed strategy implies an improvement of around 10% and 16% over the video-only DELTAS and LDA+MLLT, respectively. Part of the difference between these percentages could be associated with the effectiveness of the HiLDA technique, as we discussed in the video-only experiments.

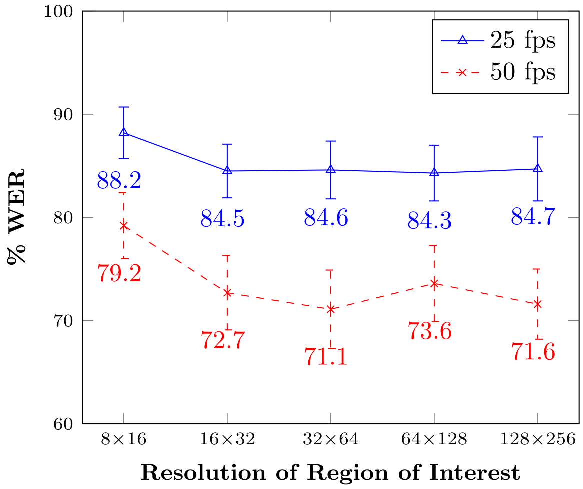

Fig. 6 Analysis of how the effectiveness (%WER) of the ViTAAl strategy (based on the audio-based alignments from the LDA+MLLT phase) is affected by the resolution of the ROI (height × width pixels) and the frame rate (frames per second (fps)). Eigenlips are used as visual speech features, since they are the best isolated representation

Regarding our visual speech feature comparison, we observe some dissimilarities w.r.t video-only experiments; for instance, eigenlips are the best isolated feature. However, similar to video-only results, the combination of features causes significant enhancements in error rates, especially when we integrate all the studied features, achieving an error rate of 68.7% WER.

Regarding the DNN-HMM system, we first estimated a visual SAT GMM-HMM model whose foundations were based on the best ViTAAl GMM-HMM system. In this way, we obtained the preferable 40-dimensional fMLLR features to train DNNs, as well as the required lattice to build the decoder of the hybrid system. Then, a DNN-HMM system was estimated using the corresponding audio-based SAT alignments. Thus, our best DNN-HMM model reached an error rate of 64.4% WER. This performance was achieved by defining a hybrid system of 2 hidden layers with 2048 neurons and parametric ReLU activation. The fMLLR features with temporal splicing of 5 context frames were used as data input. A learning rate of 0.0001 was kept during the first 20 training iterations, while the momentum and halving factor were set to a value of 0.8 and 0.4, respectively. Subsequently, by applying the SDT sMBR technique, we obtained our best VSR model capable of achieving an error rate of around 59.7% WER.

4.4 ViTAAl strategy analysis

We also analysed, as Fig. 6 reflects, how the effectiveness of our proposed ViTAAl strategy was affected by the image resolution of the ROI and the frame rate. For this experiment, different aspects were based on the results reported in Table 3. Eigenlips were used as visual speech features, since they provided (although there were no significant differences) the best recognition rate when features were compared in isolation. Regarding the ViTAAl-based system estimation, for the same reason, we used the audio alignments obtained in the LDA+MLLT training stage. Additionally, it should be noted that a different HMM topology was used when dealing with 25 frames per second (fps). Specifically, adding skip transcriptions to the final state provided the best recognition rates in this case.

As Fig. 6 shows, the frame rate stands as the main factor affecting the effectiveness of our ViTAAl strategy, observing a significant deterioration in system performance when addressing the task at 25 fps. Conversely, no significant differences exist when different image resolutions were considered, a behaviour supported by the findings reported by [38], where it was concluded that, albeit higher resolutions may be beneficial, VSR systems show a remarkable resilience to reduced resolutions.

However, we did not know the reason why the frame rate was a crucial aspect when applying ViTAAl. Hence, we performed a similar experiment to that depicted in Fig. 6 but, in this case, for the video-only scenario. The results showed that these video-only systems provided no significant differences in terms of performance regardless of image resolution and frame rate. Therefore, we were able to conclude that the quality of the acoustic alignments used to estimate our ViTAAl-based VSR system was the factor affecting the effectiveness of our strategy. In fact, it should be noted that our audio-based system achieved around 27% WER when addressing the task at 25 fps, a considerable deterioration w.r.t the 9% WER performance (see Sect. 4.1) that we were able to reach when using audio data at 50 fps.

We must be aware that this fact limits the application of our proposed ViTAAl strategy on other databases explored in the field, such as LRS2-BBC [12], LRS3-TED [52], CMU-MOSEAS [98], or LIP-RTVE [68], since they are composed of data that was recorded at 25 fps.

4.5 Overall analysis

Finally, in Table 4 we compare the best settings for each training strategy and type of system considered in our research. The first finding we can infer from these results is that, as we expected, the acoustic approach far surpasses the video-only recognition. However, by applying the ViTAAl strategy, we were able to reduce, in average terms, the error rate by 21% regarding the conventional training strategy of VSR. Concretely, our best ViTAAl setting achieved an error rate of 59.7%. These results on the VLRF corpus, as Table 5 shows, significantly improve the state of the art in the task [64, 65], highlighting that in our research we did not use data augmentation techniques.

Table 4 Overall results (%WER) depending on each training strategy and type of system proposed in our research

	Modality		
	Audio-only	**Video-only**	**ViTAAl**
GMM-HMM	9.1 ± 3.0	88.4 ± 2.6	68.7 ± 4.0
DNN-HMM	9.4 ± 3.7	84.2 ± 2.9	64.4 ± 4.0
DNN-HMM+sMBR	9.9 ± 3.4	83.4 ± 3.2	59.7 ± 4.3

Table 5 Comparison to state of the art on the VLRF task. LDA+HMMs refers to the combination of the linear discriminant analysis technique with hidden Markov models, while Conv+LSTMs refers to the combination of convolutional layers with long short-term memory networks in an end-to-end architecture. Readers are referred to [64] and [65] for a more detailed description

Method	**%WER**
LDA+HMMs [64]	80.0[a]
Conv+LSTMs end-to-end [65]	72.9[a]
ViTAAl DNN-HMM+sMBR (ours)	59.7 ± 4.3

[a] Confidence intervals were not reported

Our reports demonstrate that DNN-HMM systems, whether in combination with SDT techniques or not, improve the recognition performance of the traditional GMM-HMM paradigm for the VSR task. Nevertheless, no improvements were achieved in the audio-only scenario by using these more recent techniques. This might suggest that a lower bound has been reached in the audio-only VLRF task, a fact probably caused, as we previously mentioned in Sect. 4.1, by the scarcity of available data.

Additional experiments using state-of-the-art architectures were also conducted. By fine-tuning the pre-trained Spanish VSR model publicly released by [22] estimated with hundreds of hours, we were able to obtain recognition rates around 25% WER. However, it should be noted that not all languages have the sufficient data required to rely on these computationally expensive pre-training processes. Hence, the importance of our proposed method is that without depending on external databases it is capable of significantly improving the best results of the task to date, as reflected in Table 5. Therefore, in order to

Table 6 Decoding examples of the proposed Spanish VSR system for the VLRF corpus. Their corresponding performances are expressed in terms of word error rate (WER) and character error rate (CER). *ref* and *hyp* indicate the reference (ground truth) and the hypothesis provided by the automatic system, respectively

	Transcription	%WER	%CER
ref:	el chino vino a la escuela de intercambio	100.0	48.8
hyp:	he sido mala suerte este cambio		
ref:	tu hermano y el mio se encontraron en el metro	80.0	37.0
hyp:	su hermano enemigos encontraron espacio		
ref:	la pelicula que vimos era una comedia	57.1	21.6
hyp:	la pelicula de vemos tenia una comida		
ref:	ese ruido despertaria a todo el vecindario	42.9	14.3
hyp:	este luego despertaria todo el vecindario		
ref:	rompio una puerta de hierro	20.0	11.1
hyp:	como una puerta de hierro		
ref:	en nuestro jardin tenemos varios tipos de hierbas	12.5	8.2
hyp:	en nuestro aqui tenemos varios tipos de hierbas		
ref:	me gusta el chocolate	0.0	0.0
hyp:	me gusta el chocolate		

conduct a fair comparison and investigate to what extent these novel state-of-the-art architectures were able to deal with data scarcity scenarios when addressing the VSR task, we did not consider any pre-trained setup, i.e., we trained them from scratch similarly to what we did for ViTAAl. We first considered the so-called LF-MMI model[5], an end-to-end approach based on utterance-level discriminative training, which achieved noticeable WER reductions in multiple tasks [99, 100]. Furthermore, we also explored the architecture proposed by [22][6], one of the state-of-the-art architectures in VSR. However, non-acceptable results around 96% WER were obtained in both cases, supporting our study focused on traditional paradigms that present less dependence on large amounts of data and computational resources, aspects that can be really important when deploying applications in real scenarios.

In addition, we studied the training cost times of both paradigms in a computational resource-limited setup, namely, a 6-core 2.90 GHz Intel CPU with 16 GB memory. While estimating these end-to-end architectures following the training settings described in [22] involved around 10 h, we highlight that our proposed HMM-based model only took around 1.5 h, significantly reducing the training time cost by up to 85%.

Table 6 shows the examples of how our proposed ViTAAl-based VSR model predicted different VLRF test samples. However, although multiple qualitative and error analyses were conducted, we were not able to identify any pattern or trend in our case study. We hypothesize that the performance of a VSR system could be related to aspects that are difficult to model and that would also depend on each speaker, such as better vocalizations or certain oral physiognomies that, for some reason, reflect more adequate speech articulations.

5 Conclusions and future work

Influenced by different works studied in the field of AVSR [42, 54, 59, 62], we present the novel ViTAAl learning strategy whose purpose is to improve the performance of traditional HMM-based VSR systems by the use of auditory-based temporal alignments. We conduct experiments on the data-limited VLRF database [64], which only offers around 1 hour of training data. Results show that our proposed approach is capable of outperforming the conventional VSR training scheme in around a 24% absolute WER. In addition, we analysed how the effectiveness of the ViTAAl strategy is affected by the image resolution and the frame rate with which the visual data was collected. Our findings demonstrate that the frame rate is the main limitation of our proposed strategy, observing a significant deterioration in system performance when addressing the task at 25 fps. Conversely, as supported by [38], no significant differences exist when different image resolutions are considered. Furthermore, we conduct a comparative study on three different visual speech features based on conventional low-resource demanding techniques, concluding that the best representation is conformed by the combination of features of different natures. Thus, our research focused on the development of continuous VSR systems not only for data-scarcity scenarios, but also when computational resources are limited.

As future work, we consider exploring the ViTAAl strategy on databases that have been recorded with enough frame rates. Consequently, it would also be interesting to study the possible approaches to alleviate the limitation when addressing low frame rates. Furthermore, although different aspects should be investigated to maintain our reduced computational demand, we also consider the integration of a spelling error correction post-process using well-known large language models, such as BERT [101] or GPT4 [102].

Abbreviations

ASR	Automatic speech recognition
AVSR	Audio-visual speech recognition

[5] https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/local/chain/e2e/run_tdnn_flatstart.sh

[6] https://github.com/mpc001/Visual_Speech_Recognition_for_Multiple_Languages

CA	Convolutional autoencoders
CER	Character error rate
DNN	Deep neural networks
fMLLR	Feature space maximum likelihood linear regression
GMM	Gaussian mixture models
HMM	Hidden Markov models
LDA	Linear discriminant analysis
MFCC	Mel frequency cepstral coefficients
MLLT	Maximum likelihood linear transform
MMI	Maximum mutual information
MPE	Minimum phone error
PCA	Principal component analysis
ROI	Region of interest
SAT	Speaker adaptive training
SDT	Sequence discriminative training
sMBR	State-level minimum Bayes risk
ViTAAl	Visual training based on audio alignments
VSR	Visual speech recognition
WER	Word error rate
WFST	Weighted finite-state transducer

Acknowledgements
Not applicable.

Authors' contributions
Conceptualization: D.G.-G. and C.-D.M.-H. Methodology: D.G.-G. and C.-D.M.-H. Software: D.G.-G. Validation: D.G.-G. and C.-D.M.-H. Formal analysis: D.G.-G. and C.-D.M.-H. Investigation: D.G.-G. and C.-D.M.-H. Resources: C.-D.M.-H. Data curation: D.G.-G. and C.-D.M.-H. Writing—original draft preparation: D.G.-G. and C.-D.M.-H. Writing—review and editing: D.G.-G. and C.-D.M.-H. Visualization: D.G.-G. and C.-D.M.-H. Supervision: C.-D.M.-H. Project administration: C.-D.M.-H. Funding acquisition: C.-D.M.-H. All authors read and agreed to the submitted version of the manuscript.

Funding
This work was partially supported by Grant CIACIF/2021/295 funded by Generalitat Valenciana and by Grant PID2021-124719OB-I00 under project LLEER (PID2021-124719OB-100) funded by MCIN/AEI/10.13039/501100011033/ and by ERDF, EU A way of making Europe.

Availability of data and materials
The data sets used during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 16 July 2023 Accepted: 16 April 2024
Published online: 06 May 2024

References

1. S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimed. **2**(3), 141–151 (2000). https://doi.org/10.1109/6046.865479
2. J. Besle, A. Fort, C. Delpuech, M.-H. Giard, Bimodal speech: early suppressive visual effects in human auditory cortex. Eur. J. NeuroSci. **20**(8), 2225–2234 (2004). https://doi.org/10.1111%2Fj.1460-9568.2004.03670.x
3. H. McGurk, J. MacDonald, Hearing lips and seeing voices. Nature. **264**(5588), 746–748 (1976). https://doi.org/10.1038/264746a0
4. M. Gales, Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12** (2), 75-98 (1998). https://doi.org/10.1006/csla.1998.0043
5. B.H. Juang, L.R. Rabiner, Hidden Markov models for speech recognition. Echnometrics. **33**(3), 251–272 (1991). https://doi.org/10.2307/1268779
6. W. Chan, N. Jaitly, Q. Le, O. Vinyals, *ICASSP*. Listen, attend and spell: a neural network for large vocabulary conversational speech recognition (2016), pp. 4960-4964
7. P. Ma, S. Petridis, M. Pantic, *ICASSP*. End-to-end audio-visual speech recognition with conformers (IEEE, 2021), pp. 7613–7617
8. A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision (2022). arXiv preprint arXiv:2212.04356
9. M. Anwar, B. Shi, V. Goswami, W. Hsu, J. Pino, C. Wang, *Interspeech*. MuAViC: a multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation (ISCA, 2023), pp. 4064–4068
10. M. Burchi, R. Timofte, *Wacv*. Audio-visual efficient conformer for robust speech recognition (2023), pp. 2257-2266
11. B. Juang, Speech recognition in adverse environments. Comput. Speech Lang. **5**(3), 275–294 (1991). https://doi.org/10.1016/0885-2308(91)90011-E
12. T. Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep audio-visual speech recognition. Trans. PAMI. (2018). https://doi.org/10.1109/TPAMI.2018.288905
13. G. Potamianos, C. Neti, G. Gravier, A. Garg, A. Senior, Recent advances in the automatic recognition of audiovisual speech. IEEE. **91**(9), 1306–1326 (2003). https://doi.org/10.1109/JPROC.2003.817150
14. B. Shi, W.N. Hsu, K. Lakhotia, A. Mohamed, Learning audio-visual speech representation by masked multimodal cluster prediction (2022). arXiv preprint arXiv:2201.02184
15. P. Eickhoff, M. Möller, T.P. Rosin, J. Twiefel, S. Wermter, *ICANN* (Introducing Noise Robustness to Pretrained Automatic Speech Recognition (Springer, Nature Switzerland, Bring the Noise, 2023)
16. Z. Huang, S. Watanabe, S.-W. Yang, P. García, S. Khudanpur, *ICASSP*. Investigating self-supervised learning for speech enhancement and separation (2022), pp. 6837-6841
17. S. Pascual, A. Bonafonte, J. Serrá, *Interspeech*. SEGAN: speech enhancement generative adversarial network (ISCA, 2017), pp. 3642–3646
18. H. Yen, F. Germain, G. Wichern, J. Roux, *ICASSP*. Cold diffusion for speech enhancement (IEEE, 2023), pp. 1-5
19. A. Fernandez-Lopez, F.M. Sukno, Survey on automatic lip-reading in the era of deep learning. Image Vision Comput. **78**, 53–72 (2018). https://doi.org/10.1016/j.imavis.2018.07.002
20. A. Fernandez-Lopez, F.M. Sukno, *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Optimizing phoneme-to-viseme mapping for continuous lip-reading in spanish (Elsevier, 2017), pp. 305–328
21. K. Thangthai, Computer lipreading via hybrid deep neural network hidden Markov models (Unpublished doctoral dissertation) (University of East Anglia, 2018)
22. P. Ma, S. Petridis, M. Pantic. Visual speech recognition for multiple languages in the wild. Nat. Mach. Intel. **4**(11), 930–939 (2022). https://doi.org/10.1038/s42256-022-00550-z
23. M. Ezz, A.M. Mostafa, A.A. Nasr, A silent password recognition framework based on lip analysis. IEEE Access **8**, 55354–55371 (2020). https://doi.org/10.1109/ACCESS.2020.2982359
24. T. Stafylakis, G. Tzimiropoulos, *ECCV*. Zero-shot keyword spotting for visual speech recognition in-the-wild (2018), pp. 513–529
25. B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, J.S. Brumberg, Silent speech interfaces. Speech Commun. **52**(4), 270–287 (2010). https://doi.org/10.1016/j.specom.2009.08.002
26. J.A. Gonzalez-Lopez, A. Gomez-Alanis, J.M. Martín Doñas, J.L. Pérez-Córdoba, A.M. Gomez, Silent speech interfaces for speech restoration: a review. IEEE Access. **8**, 177995–178021 (2020). https://doi.org/10.1109/ACCESS.2020.3026579
27. G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Proc. Mag. **29**(6), 82–97 (2012). https://doi.org/10.1109/MSP.2012.2205597
28. K. Veselý, A. Ghoshal, L. Burget, D. Povey, *Interspeech*. Sequence-discriminative training of deep neural networks (2013), pp. 2345–2349

29. R. Prabhavalkar, T. Hori, T. Sainath, R. Schlüter, S. Watanabe, End-to-end speech recognition: a survey (2023). arXiv preprint arXiv:2303.03329
30. M. Gales, S. Young, *The application of hidden Markov models in speech recognition* (Now Publishers Inc., Now Foundations and Trends, 2008)
31. C. Fisher, Confusions among visually perceived consonants. J. Speech Hear. Res. **11**(4), 796–804 (1968). https://doi.org/10.1044/jshr.1104.796
32. H. Bear, R. Harvey, B. Theobald, Y. Lan, *International Symposium on Visual Computing*. Which phoneme-to-viseme maps best improve visual-only computer lip-reading? (Springer, 2014), pp. 230–239
33. L. Cappelletta, N. Harte, *19th European Signal Processing Conference*. Viseme definitions comparison for visual-only speech recognition (2011), pp. 2109-2113
34. D. Howell, S. Cox, B. Theobald, Visual units and confusion modelling for automatic lip-reading. Image Vision Comput. **51**, 1–12 (2016). https://doi.org/10.1016/j.imavis.2016.03.003
35. K. Thangthai, R. Harvey, *Interspeech*. Building large-vocabulary speaker-independent lipreading systems (ISCA, 2018), pp. 2648–2652
36. K. Thangthai, R. Harvey, S. Cox, B. Theobald, *AVSP*. Improving lip-reading performance for robust audiovisual speech recognition using DNNs (2015), pp. 127–131
37. H. Bear, R. Harvey, *ICASSP*. Decoding visemes: improving machine lip-reading (2016), pp.2009–2013
38. H. Bear, R. Harvey, B. Theobald, Y. Lan, *ICIP*. Resolution limits on visual speech recognition (IEEE, 2014), pp. 1371–1375
39. I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading. IEEE Trans. PAMI **24**(2), 198–213 (2002). https://doi.org/10.1109/34.982900
40. A.A. Shaikh, D.K. Kumar, W.C. Yau, C. Azemin, J. Gubbi, 3rd CISP. Lip reading using optical flow and support vector machines. IEEE. **1**, 327–330 (2010)
41. D. Parekh, A. Gupta, S. Chhatpar, A. Yash, M. Kulkarni, *5th I2CT*. Lip reading using convolutional auto encoders as feature extractor (2019), pp. 1–6
42. P. Ma, R. Mira, S. Petridis, B.W. Schuller, M. Pantic, *Interspeech*. LiRA: learning visual speech representations from audio through self-supervision (2021), pp.3011–3015
43. P. Duchnowski, D.S. Lum, J.C. Krause, M.G. Sexton, M.S. Bratakos, L.D. Braida, Development of speechreading supplements based on automatic speech recognition. IEEE rans. Biomed. Eng. **47**(4), 487–496 (2000). https://doi.org/10.1109/10.828148
44. Y. Lan, R. Harvey, B. Theobald, E. Ong, R. Bowden, *International Conference on Auditory-Visual Speech Processing*. Comparing visual features for lipreading (2009), pp. 102–106
45. K. Thangthai, R. Harvey, *Interspeech*. Improving computer lipreading via DNN sequence discriminative training techniques (2017), pp. 3657–3661
46. N. Harte, E. Gillen, TCD-TIMIT: an audio-visual corpus of continuous speech. IEEE Trans. Multimed. **17**(5), 603–615 (2015). https://doi.org/10.1109/TMM.2015.2407694
47. K. Thangthai, H. Bear, R. Harvey, *BMVC*. Comparing phonemes and visemes with DNN-based lipreading (2017), pp. 4–7
48. P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, M. Pantic, *ICASSP*. Auto-AVSR: audio-visual speech recognition with automatic labels (2023), pp.1–5
49. H. Chen, H. Zhou, J. Du, C.-H. Lee, J. Chen, S. Watanabe, C. Liu, *ICASSP*. The first multimodal information based speech processing (Misp) challenge: data, tasks, baselines and results (IEEE, 2022), pp. 9266-9270
50. K.R. Prajwal, T. Afouras, A. Zisserman, *CVPR*. Sub-word level lip reading with visual attention (IEEE, 2022), pp. 5162-5172
51. J. Son Chung, A. Senior, O. Vinyals, A. Zisserman, *CVPR*. Lip reading sentences in the wild (2017), pp. 6447–6456
52. T. Afouras, J.-S. Chung, A. Zisserman, LRS3-TED: a large-scale dataset for visual speech recognition. (2018). arXiv preprint arXiv:1809.00496
53. S. Bhati, J. Villalba, L. Moro-Velazquez, T. Thebaud, N. Dehak, Leveraging pretrained image-text models for improving audio-visual learning (2023). arXiv preprint arXiv:2309.04628
54. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Ng, *28th ICML*. Multimodal deep learning (PMLR, 2011), pp. 689–696
55. A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, I. Sutskever, *ICML*. Learning transferable visual models from natural language supervision, vol. 139 (PMLR, 2021), pp. 8748–8763
56. E. Petajan, *CVPR*. Automatic lipreading to enhance speech recognition (IEEE, 1985), pp. 40–47
57. A. Adjoudani, C. Benoît, *Speechreading by humans and machines*. On the integration of auditory and visual parameters in an HMM-based ASR. (Springer, 1996), pp. 461–471
58. P. Teissier, J. Robert-Ribes, J. Schwartz, A. Guérin-Dugué, Comparing models for audiovisual fusion in a noisy-vowel recognition task. IEEE Trans. Speech Audio Process. **7**(6), 629–642 (1999). https://doi.org/10.1109/89.799688
59. T. Afouras, J.S. Chung, A. Zisserman, *ICASSP*. ASR is all you need: cross-modal distillation for lip reading (2020), pp.2143–2147
60. Y.A.D. Djilali, S. Narayan, H. Boussaid, E. Almazrouei, M. Debbah, *ICCV*. Lip2Vec: efficient and robust visual speech recognition via Latentto-Latent Visual to Audio Representation Mapping (IEEE, 2023), pp. 13790-13801
61. C. Sui, M. Bennamoun, R. Togneri, *ICCV*. Listening with your eyes: towards a practical visual speech recognition system using deep Boltzmann machines (2015), pp. 154–162
62. A. Thanda, S. Venkatesan, Multi-task learning of deep neural networks for audio visual automatic speech recognition (2017). arXiv preprint arXiv:1701.02477
63. R. Caruana, Multitask learning. Mach. Learn. **28**(1), 41–75 (1997). https://doi.org/10.1023/A:1007379606734
64. A. Fernandez-Lopez, O. Martinez, F.M. Sukno, *12th FG*. Towards estimating the upper bound of visual-speech recognition: the visual lip-reading feasibility database (2017), pp.208–215
65. A. Fernandez-Lopez, F. Sukno, End-to-end lip-reading without large-scale data. IEEE/ACM TASLP. **30**, 2076–2090 (2022). https://doi.org/10.1109/TASLP.2022.3182274
66. D. Gimeno-Gomez, C.-D. Martinez-Hinarejos, *IberSPEECH*. Speaker-adapted endto-end visual speech recognition for continuous Spanish (2022), pp. 41–45
67. D. Gimeno-Gómez, C.-D. Martínez-Hinarejos, *IberSPEECH*. Analysis of visual features for continuous lipreading in Spanish (2021), pp. 220–224
68. D. Gimeno-Gómez, C.-D. Martínez-Hinarejos, *LREC*. LIP-RTVE: an audiovisual database for continuous Spanish in the wild (ELRA, 2022), pp.2750–2758
69. N. Ahmed, T. Natarajan, K. Rao, Discrete cosine transform. IEEE Trans. Comput. **100**(1), 90–93 (1974). https://doi.org/10.1109/T-C.1974.223784
70. D. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
71. P. Wiggers, J.C. Wojdel, L. Rothkrantz, *7th ICSLP*. Medium vocabulary continuous audio-visual speech recognition (ISCA, 2002), pp. 1921–1924
72. G. Bradski, The opencv library. Dr Dobb's J. Softw. Tools. **25**, 120–125 (2000)
73. D. King, Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
74. V. Kazemi, J. Sullivan, *CVPR*. One millisecond face alignment with an ensemble of regression trees (2014), pp. 1867–1874
75. O. Koller, J. Forster, H. Ney, Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. Comput. Vision Image Underst. **141**, 108–125 (2015). https://doi.org/10.1016/j.cviu.2015.09.013
76. A. Chitu, L. Rothkrantz, Visual speech recognition automatic system for lip reading of Dutch. J. Inf. Technol. Control. **3**, 2–9 (2009). https://doi.org/10.5772/36466
77. K. Delac, M. Grgic, P. Liatsis, *47th ELMAR*. Appearance-based statistical methods for face recognition (IEEE, 2005), pp. 151–158
78. S. Wold, K. Esbensen, P. Geladi, Principal component analysis. Chemometr. Intell. Lab. Syst. **2**(1–3), 37–52 (1987). https://doi.org/10.1016/0169-7439(87)80084-9
79. I. Fung, B. Mak, *IEEE ICASSP*. End-to-end low-resource lip-reading with maxout CNN and LSTM. (IEEE, 2018), pp. 2511–2515
80. K. Paleček, *International Conference on Speech and Computer*. Extraction of features for lip-reading using autoencoders (2014), pp. 209–216
81. Y. Bengio, Learning deep architectures for AI. Found. Trends Mach. Learn. **2**(1), 1–127 (2009). https://doi.org/10.1561/2200000006
82. G. Potamianos, J. Luettin, C. Neti, *ICASSP*. Hierarchical discriminant features for audio-visual LVCSR, vol. 1 (2001), pp. 165–168

83. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, K. Vesely, *ASRU*. The Kaldi Speech Recognition Toolkit (IEEE Signal Processing Society, 2011)
84. C. Rao, *Linear statistical inference and is applications* (John Wiley & Sons, New York, 1965)
85. R. Gopinath, ICASSP. Maximum likelihood modeling with Gaussian distributions for classification **2**, 661–664 (1998)
86. T. Anastasakos, J. McDonough, J. Makhoul, ICASSP. Speaker adaptive training: a maximum likelihood approach to speaker normalization. IEEE. **2**, 1043–1046 (1997)
87. G.E. Hinton, in *Neural Networks: Tricks of the Trade: Second Edition*. A practical guide to training restricted Boltzmann machines (Berlin, Heidelberg, Springer Berlin Heidelberg, 2012), pp. 599–619
88. B. Kingsbury, *ICASSP*. Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling (IEEE, 2009), pp. 3761-3764
89. G. Wang, K.C. Sim, *Interspeech*. Sequential classification criteria for NNs in automatic speech recognition (ISCA, 2011), pp. 441–444
90. L. Bahl, P. Brown, P. de Souza, R. Mercer, ICASSP. Maximum mutual information estimation of hidden Markov model parameters for speech recognition **11**, 49–52 (1986)
91. D. Povey, P. Woodland, *ICASSP*. Minimum phone error and I-smoothing for improved discriminative training, vol. 1 (2002), pp. I-105-I-108
92. J. Kaiser, B. Horvat, Z. Kacic, ICSLP. A novel loss function for the overall risk criterion based discriminative training of HMM models **2**, 887–890 (2000)
93. D. Povey, B. Kingsbury, *ICASSP*. Evaluation of proposed modifications to MPE for large scale discriminative training, vol 4 (IEEE, 2007), pp. IV-321-IV-324
94. M. Mohri, F. Pereira, M. Riley, *Springer Handbook of Speech Processing*. Speech recognition with weighted finite-state transducers (Springer, 2008), pp. 559–584
95. A. Quilis, Principios de fonología y fonética españolas, vol. 43 (Arco Libros, 1997)
96. A. Stolcke, *ICSLP*. SRILM – an extensible language modeling toolkit (ISCA, 2002), pp. 901–904
97. M. Bisani, H. Ney, ICASSP. Bootstrap estimates for confidence intervals in ASR performance evaluation. IEEE. **1**, 409–412 (2004)
98. A. Zadeh, Y. Cao, S. Hessner, P. Liang, S. Poria, L. Morency, *EMNLP*. MOSEAS: a multimodal language dataset for Spanish, Portuguese, German and French (ACL, 2020), pp. 1801–1812
99. H. Hadian, H. Sameti, D. Povey, S. Khudanpur, *Interspeech*. End-to-end speech recognition using lattice-free MMI (ISCA, 2018), pp. 12–16
100. H. Hadian, H. Sameti, D. Povey, S. Khudanpur, Flat-start single-stage discriminatively trained HMM-based models for ASR. IEEE/ACM TASLP. **26**(11), 1949–1961 (2018). https://doi.org/10.1109/TASLP.2018.2848701
101. O. Hrinchuk, M. Popova, B. Ginsburg, *ICASSP*. Correction of automatic speech recognition with transformer sequence-to-sequence model (IEEE, 2020), pp. 7074–7078
102. L. Mai, J. Carson-Berndsen, Enhancing conversational quality in language learning chatbots: an evaluation of GPT4 for ASR error correction (2023). arXiv preprint arXiv:2307.09744

Publisher’s Note