

Learning of Meaningful Visual Representations for Continuous Lip-Reading

Adriana Fernández López

TESI DOCTORAL UPF / Year 2020

THESIS SUPERVISOR

Federico M. Sukno

Department of Information and Communications Technologies



you do not just wake up
and become the butterfly
growth is a process

rupi kaur

i stand
on the sacrifices
of a million women before me
thinking
what can i do
to make this mountain taller
so the women after me
can see farther

legacy - rupi kaur

Acknowledgments

This is my opportunity to say thank you to some wonderful people without whom I really could not have finished my PhD.

First of all, I would like to thank my supervisor, Professor Federico M. Sukno, for your support and guidance in the whole process. You believed straight forward in my career and the potential of the topic since my very beginnings, so thank you for pushing me beyond I would have never done. I will take the nice experiences with me.

I would like to acknowledge the colleagues from my research stay at Trinity College Dublin for their wonderful collaboration. They made everything very easy and I felt at home. I would especially like to single out my supervisor at TCD Professor Naomi Harte. Naomi, I want to thank you for your positivism and support and for all the opportunities I was given to further my research. This experience will always fly with me.

I would also like to thank all my colleagues from UPF, who made the time during my PhD study worth it, thank you for making coming in to work every day good. Especially, Hermann who always was there supporting me. I would also like to thank all the administrative staff at the DTIC.

També vull agrair a tots els voluntaris que van participar en la base de dades VLRF. Gràcies pel vostre temps i paciència. La vostra col·laboració ha estat molt important en el nostre projecte. Especialment, vull agrair la col·laboració d'ACCAPS, la Federació d'Associacions Catalanes de Pares i Persones Sordes, per la seva implicació durant aquest projecte. Pas a pas, ajudarem a millorar la qualitat de vida i la integració de les persones amb discapacitat auditiva a la societat.

Por último, me gustaría agradecer a mi familia y amigos. Primero, me gustaría agradecer a mis padres y mi hermano, quienes me han apoyado incondicionalmente durante este loco proceso. Gracias por estar siempre ahí para mí. También me gustaría agradecer especialmente a mis mejores amigas, sin vosotras este proceso hubiera sido mucho más duro. Elena, Esther, Mika, Noelia, Sara, Sonia, Yasmina, María, Indira y ambas Adrianas, gracias por escalar la montaña rusa conmigo y enseñarme a verme como me veis vosotras, con los mejores ojos que podría tener. Muchas gracias, os adoro.

Abstract

In the last decades, there has been an increased interest in decoding speech exclusively using visual cues, i.e. mimicking the human capability to perform lip-reading, leading to Automatic Lip-Reading (ALR) systems. However, it is well known that the access to speech through the visual channel is subject to many limitations when compared to the audio channel, i.e. it has been argued that humans can actually read around 30% of the information from the lips, and the rest is filled-in from the context. Thus, one of the main challenges in ALR resides in the visual ambiguities that arise at the word level, highlighting that not all sounds that we hear can be easily distinguished by observing the lips.

In the literature, early ALR systems addressed simple recognition tasks such as alphabet or digit recognition but progressively shifted to more complex and realistic settings leading to several recent systems that target continuous lip-reading. To a large extent, these advances have been possible thanks to the construction of powerful systems based on deep learning architectures that have quickly started to replace traditional systems. Despite the recognition rates for continuous lip-reading may appear modest in comparison to those achieved by audio-based systems, the field has undeniably made a step forward. Interestingly, an analogous effect can be observed when humans try to decode speech: given sufficiently clean signals, most people can effortlessly decode the audio channel but would struggle to perform lip-reading, since the ambiguity of the visual cues makes it necessary the use of further context to decode the message.

In this thesis, we explore the appropriate modeling of visual representations with the aim to improve continuous lip-reading. To this end, we present different data-driven mechanisms to handle the main challenges in lip-reading related to the ambiguities or the speaker dependency of visual cues.

Our results highlight the benefits of a proper encoding of the visual channel, for which the most useful features are those that encode corresponding lip positions in a similar way, independently of the speaker. This fact opens the door to i) lip-reading in many different languages without requiring large-scale datasets, and ii) increasing the contribution of the visual channel in audio-visual speech systems. On the other hand, our experiments identify a tendency to focus on the modeling of temporal context as the key to advance the field, where there is a need for ALR models that are trained on datasets comprising large speech variability at several context levels. In this thesis, we show that both proper modeling of visual representations and the ability to retain context at several levels are necessary conditions to build successful lip-reading systems.

Resum

En les darreres dècades, hi ha hagut un interès creixent en la descodificació de la parla utilitzant exclusivament senyals visuals, és a dir, imitant la capacitat humana de llegir els llavis, donant lloc a sistemes de lectura automàtica de llavis (ALR). No obstant això, se sap que l'accés a la parla a través del canal visual està subjecte a moltes limitacions en comparació amb el senyal acústic, és a dir, s'ha argumentat que els humans poden llegir al voltant del 30% de la informació dels llavis, i la resta es completa fent servir el context. Així, un dels principals reptes de l'ALR resideix en les ambigüitats visuals que sorgeixen a escala de paraula, destacant que no tots els sons que escoltem es poden distingir fàcilment observant els llavis.

A la literatura, els primers sistemes ALR van abordar tasques de reconeixement senzilles, com ara el reconeixement de l'alfabet o els dígitos, però progressivament van passar a entorns més complexos i realistes que han conduït a diversos sistemes recents dirigits a la lectura contínua dels llavis. En gran manera, aquests avenços han estat possibles gràcies a la construcció de sistemes potents basats en arquitectures d'aprenentatge profund que han començat a substituir ràpidament els sistemes tradicionals. Tot i que les taxes de reconeixement de la lectura contínua dels llavis poden semblar modestes en comparació amb les assolides pels sistemes basats en àudio, és evident que el camp ha fet un pas endavant. Curiosament, es pot observar un efecte anàleg quan els humans intenten descodificar la parla: donats senyals sense soroll, la majoria de la gent pot descodificar el canal d'àudio sense esforç, però tindria dificultats per llegir els llavis, ja que l'ambigüitat dels senyals visuals fa necessari l'ús de context addicional per descodificar el missatge.

En aquesta tesi explorem el modelatge adequat de representacions visuals amb l'objectiu de millorar la lectura contínua dels llavis. Amb aquest objectiu, presentem diferents mecanismes basats en dades per fer front als principals reptes de la lectura de llavis relacionats amb les ambigüitats o la dependència dels parlants dels senyals visuals.

Els nostres resultats destaquen els avantatges d'una correcta codificació del canal visual, per a la qual les característiques més útils són aquelles que codifiquen les posicions corresponents dels llavis d'una manera similar, independentment de l'orador. Aquest fet obre la porta a i) la lectura de llavis en molts idiomes diferents sense necessitat de conjunts de dades a gran escala, i ii) a l'augment de la contribució del canal visual en sistemes de parla audiovisuals. D'altra banda, els nostres experiments identifiquen una tendència a centrar-se en la modelització del context temporal com la clau per avançar en el camp, on hi ha la necessitat de models d'ALR que s'entrenin en conjunts de dades que incloguin una gran variabilitat de la parla a diversos nivells de context. En aquesta tesi,

demostrarem que tant el modelatge adequat de les representacions visuals com la capacitat de retenir el context a diversos nivells són condicions necessàries per construir sistemes de lectura de llavis amb èxit.

Contents

List of figures	xvii
List of tables	xx
List of Abbreviations	xxi
1 INTRODUCTION	1
1.1 Contribution	4
1.2 Significance	5
1.2.1 Speech Recognition Systems	6
1.2.2 Hearing-impaired people	6
1.3 Outline of the thesis	7
1.4 Publications from this thesis	8
2 REVIEW OF THE LITERATURE	11
2.1 Audio-Visual databases	12
2.1.1 Alphabet and digit recognition	13
2.1.2 Word and sentence recognition	16
2.1.3 Multiview databases	19
2.2 Automatic lip-reading systems	21
2.2.1 Traditional ALR systems	22
2.2.2 DNN-based ALR systems	28
2.3 Summary and Conclusions	44
3 OPTIMIZING PHONEME-TO-VISEME MAPPING FOR CONTINUOUS LIP-READING IN SPANISH	47
3.1 ALR system	50
3.1.1 Related Work	50
3.1.2 Our System	51
3.2 Experiments	56
3.2.1 Databases	56

3.2.2	Phonetic alphabet	58
3.2.3	Results	58
3.3	Discussion	61
3.4	Conclusions	67
4	THE UPPER BOUND OF VISUAL SPEECH RECOGNITION	69
4.1	Audio-visual speech databases	71
4.2	Visual Lip-Reading Feasibility Database	72
4.2.1	Participants	73
4.2.2	Utterances	74
4.2.3	Technical aspects	74
4.2.4	Data labeling	75
4.3	Results	76
4.3.1	Experimental setup	76
4.3.2	Human lip-reading	77
4.3.3	Training and context influence on lip-reading	78
4.3.4	Human observers and automatic system comparison	79
4.4	Discussion and Conclusions	82
5	END-TO-END LIP-READING WITHOUT LARGE-SCALE DATA	87
5.1	Training with limited data	90
5.1.1	Visual Units	91
5.2	Frame mapping into visual units	98
5.2.1	Deep latent features	98
5.2.2	Optimal number of visual units per subject	100
5.2.3	Generalization of all speaker-specific sets into a common set	101
5.3	Spatio-temporal data augmentation	102
5.3.1	Motivation	102
5.3.2	Preprocessing the data	104
5.3.3	Synthesis of video sequences	105
5.4	Proposed Lip-Reading Architecture	108
5.4.1	Input pre-processing	110
5.4.2	Visual module	111
5.4.3	Temporal module	111
5.5	Experiments in a constrained scenario	112
5.5.1	Database	113
5.5.2	Results	113
5.5.3	Conclusions	115
5.6	Experiments in continuous speech	115
5.6.1	The VLRF dataset	115

5.6.2	External language model	116
5.6.3	Results	116
5.6.4	Temporal module	119
5.6.5	Discussion and Conclusions	122
6	VISUAL SPEECH ADAPTATION TO NEW SPEAKERS	125
6.1	Speaker Adaptation	126
6.2	Proposed AVSR System	127
6.2.1	CoGANs	127
6.2.2	Proposed architecture	128
6.3	Experiments and Results	130
6.3.1	The TCD-TIMIT dataset	130
6.3.2	Training procedure	131
6.3.3	Learning a joint distribution	131
6.3.4	Comparison between speech recognition systems	132
6.4	Discussion and Conclusions	133
7	CONCLUSIONS	135
7.1	Research summary	135
7.2	Discussion and future work	137

List of Figures

2.1	Cumulative number of papers on ALR systems published between 2007 and 2017.	12
2.2	Example shots of audio-visual speech databases.	13
2.3	The main processing blocks of an ALR system	21
2.4	Digit and alphabet recognition. Left-side: number of times that each feature technique has been used from 2007 to 2017; Right-side: number of times that each classification method has been used from 2007 to 2017.	23
2.5	Cumulative number of ALR systems targeting <i>digit or alphabet</i> and <i>word or sentence</i> recognition from 2007 to 2017.	25
2.6	Word and sentence recognition. Number of times that each feature technique has been used from 2007 to 2017.	26
2.7	Word and sentence recognition. Number of times that each classification method has been used from 2007 to 2017.	27
2.8	DNN-based systems. Number of times that each feature technique has been used from 2007 to 2018.	33
2.9	DNN-based systems. Number of times that each classification method has been used from 2007 to 2018.	33
2.10	Baseline DL architecture for lip-reading, consisting of combinations of CNNs and LSTMs.	34
2.11	Architectures from [43]. (a) Combination of SyncNet and LSTMs; (b) Combination of VGG-M and LSTMs	35
2.12	(a) Architecture from [117]; (b) Architecture from [11]	35
2.13	Architecture from [205]	36
2.14	Architecture from [44] (a) WLAS; (b) WATCH; (c) SPELL	37
2.15	(a) Architecture from [228]; (b) Architecture from [229]	39
2.16	(a) Architecture from [42]; (b) Architecture from [172]	40
3.1	General process of an ALR system.	52
3.2	(a) IOF-ASM detection, the marks in yellow are used to fix the bounding box; (b) ROI detection, each color fix a lateral of the bounding box; (c) Keypoints distribution.	52

3.3	(a) Probability density functions for in-class (green) and out-of-class (red) samples; (b) Cumulative distributions corresponding to (a). Notice than for in-class samples we use the complement of the cumulative distribution, since lower values should have higher probabilities. Reprinted from [63].	55
3.4	Comparison of features performance. Reprinted from [63]	59
3.5	Boxplots of system performance the AV@CAR database in terms of viseme-, phoneme- and word accuracy for different vocabularies. We analyze the one-to-one mapping phoneme-to-viseme, and the many-to-one phoneme-to-viseme mappings with 23, 20, 16 and 14 visemes. The phoneme accuracy is always computed from the 28 phonemes.	60
3.6	Comparison of system performance in the AV@CAR database in terms of word accuracy for the different vocabularies and participants.	62
3.7	(a) Resulting confusion matrix from a system trained in VLRf using 20 visemes (many-to-one phoneme-to-viseme mapping). (b) Resulting confusion matrix from a system trained in VLRf using 28 visemes (one-to-one phoneme-to-viseme mapping). Additionally, we highlighted in yellow, the phonemes that share the same viseme in the proposed alphabet to a clearer comprehension.	65
3.8	Frequency of appearance of each phoneme in the VLRf database.	66
4.1	Scheme of the recording setup and snapshots of the VLRf database.	75
4.2	Word accuracy for normal-hearing (H) and hearing-impaired groups (H-Imp) at each repetition.	77
4.3	Word accuracy per participant at each repetition.	77
4.4	Word recognition average for each participant at each level.	79
4.5	Cumulative average per sentence for all participants at each repetition.	80
4.6	Top: system performance in terms of word recognition rate for each participant. Bottom: system performance in terms of phoneme recognition rate for each participant.	81
4.7	Top: human observers performance (Repetition 1) and automatic system performance for each participant in terms of word recognition average; Bottom: human observers performance (Repetition 1) and automatic system performance for each participant in terms of phoneme recognition average.	82

4.8	Top: Number of wrong detected phonemes. The red columns represent the false negatives phonemes and the green ones the false positives.; Bottom: Precision and Recall of each phoneme.	83
5.1	Mapping into visual units of the phrase "Miraba el reloj" for sets with 15 and 11 visual units.	95
5.2	Example of MAD for frames $m = 60$ and $m = 80$ for the phrase "Miraba el reloj".	100
5.3	Example of approximate annotations per frame for 3 sequences.	102
5.4	Collecting the phonetic dataset per subject	105
5.5	Interpolation example.	107
5.6	Wrong interpolation examples.	107
5.7	Proposed ALR system.	109
5.8	Confusion matrices: a) Phoneme labels from VLRf; b) Visual units derived using phoneme labels from VLRf. The phonemes can be found in the following order :/o/, /m/, /k/, /w/, /t/, /jj/, /l/, /x/, /L/, /u/, /g/, /z/, /d/, /G/, /A/, /r/, /T/, /b/, /j/, /s/, /e/, /p/, /n/, /N/, /J/, /B/, /D/, /i/, /tS/, /a/, /f/.	117
5.9	Confusion matrices: c) Estimated phoneme labels; d) Visual units derived using estimated phoneme labels. The phonemes can be found in the following order :/o/, /m/, /k/, /w/, /t/, /jj/, /l/, /x/, /L/, /u/, /g/, /z/, /d/, /G/, /A/, /r/, /T/, /b/, /j/, /s/, /e/, /p/, /n/, /N/, /J/, /B/, /D/, /i/, /tS/, /a/, /f/.	119
5.10	A visualization of attention weights evaluated on system \mathcal{B} : a) Baseline; b) Baseline + VU; c) Baseline + DAS; d) Baseline + DAS + VU. We only observe soft-alignment between source video sequence and target characters for Baseline + DAS + VU.	122
5.11	CER and WER for different ALR systems conditioned to the number of training sentences.	124
6.1	Proposed Speaker Adapted-AVSR system. On the left we see the feature adaptation using CoGANs: On the right, the AVSR system. Both networks are jointly trained to adapt the visual front-end to a new speaker.	129
6.2	Synthesized images for different z when: (a) $D^1 \equiv D^2$; (b) $\theta_{D^1_{(1)}} \neq \theta_{D^2_{(1)}}$ and $\theta_{D^1_{(l)}} = \theta_{D^2_{(l)}}$, for $l = 2, 3, \dots, L_D$	132
6.3	(a) CER on SI partition of TCD-TIMIT and (b) Adapted-AVSR with reduced P^2	133

List of Tables

2.1	Audio-visual corpora, in chronological order	15
2.2	Sentence examples of audio-visual databases	16
2.3	Multi-view audio-visual databases, in chronological order	19
2.4	ALR systems from 2007 to 2017 - Part I	29
2.5	ALR systems from 2007 to 2017 - Part II	30
2.6	ALR systems from 2007 to 2017 - Part III	31
3.1	Sample sentences for each database and their corresponding phonetic transcription using SAMPA.	57
3.2	Average lip-images per user and phoneme of 5 subjects of the AV@CAR database. Each row shows a sample subject. For each subject, every column shows the average of all the frames in which the subject uttered a specific phoneme. The vertical lines separate the phonemes that belong to different visemes according to our mapping. The last row shows the average when considering all the users together.	63
3.3	Average lip-images per user and phoneme of 5 subjects of the VLRf database. Each row shows a sample subject. For each subject, every column shows the average of all the frames in which the subject uttered a specific phoneme. The vertical lines separate the phonemes that belong to different visemes according to our mapping. The last row shows the average when considering all the users together.	64
3.4	System performance in the VLRf database in terms of viseme-, phoneme- and word accuracy for the vocabularies of 20 and 28 classes.	65
4.1	Statistical comparison between hearing-impaired and normal-hearing participants at each repetition.	78
5.1	Some of the largest audio-visual databases for continuous speech recognition for various languages.	88

5.2	Visual module details	110
5.3	Comparison with previous work on the OuluVS2 database.	115
5.4	Character Error Rate (CER) and Word Error Rate (WER) for ALR systems \mathcal{A} and \mathcal{B} evaluated on the VLR dataset for different experimental conditions. We show our results when using Greedy (G) decoding, Beam Search (BS) decoding and an additional external language model (LM) with BS.	120
5.5	Examples of ALR results. GT : Ground Truth; G : Greedy, BS : Beam Search; and LM : Language Model.	123
6.1	GANs architecture for speaker adaptation	130

List of Abbreviations

Abbreviation	Meaning
AAM	Active Appearance Model
ALR	Automatic Lip-Reading
ASM	Active Shape Model
ASR	Automatic Speech Recognition
AVSR	Audio-Visual Speech Recognition
CD	Context Dependent
CER	Character Error Rate
CFI	Concatenated Frame Image
CHAVF	Cascade Hybrid Appearance Visual Feature
CI	Context Independent
CNN	Convolutional Neural Network
CoGAN	Coupled Generative Adverarial Network
CR	Classification Rate
CTC	Connectionist Temporal Classification
DA	Data Augmentation
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DL	Deep Learning
DNN	Deep Neural Network
DWT	Discrete Wavelet Transform
FC	Fully Connected
FDCT	Fast Discrete Curvelet Transform
GAN	Generative Adverarial Network
GRU	Gated Recurrent Unit
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
IOF	Invariant Optimal Features
KD	Knowledge Distillation
K-NN	K- Nearest Neighbors
LBP	Local Binary Pattern

Abbreviation	Meaning
LDA	Linear Discriminant Analysis
LDG	Locality Discriminant Graph
LSTM	Long-Short Term Memory
LV	Latent Variable
MBH	Motion Boundary Histograms
MFCC	Mel-Frequency Cepstral Coefficients
MLLT	Maximum Likelihood Linear Transform
MLP	Multi-Layer Perceptron
NIN	Networks in Networks
PCA	Principal Component Analysis
PLVM	Proposed Latent Variable Model
RBM	Restricted Boltzmann Machines
RDA	Regularized Discriminant Analysis
RFMA	Random Forest Manifold Alignment
RNN	Recurrent Neural Network
ROI	Region of Interest
SDF	Shape Difference Feature
SDM	Supervised Descend Method
SIFT	Scale-Invariant Feature Transform
SP	Sequential Pattern
SNR	Signal to Noise Ratio
STLF	Spatio-Temporal Lip Feature
SVM	Support Vector Machine
TCN	Temporal Convolutional Network
TGD	Temporal Gradient Descend
VSR	Visual Speech Recognition
WER	Word Error Rate
WRR	Word Recognition Rates

Chapter 1

INTRODUCTION

Speech is the most used communication method between humans, and it is considered a multi-sensory process that involves the perception of both acoustic and visual cues. McGurk and MacDonald demonstrated the influence of vision in speech perception in 1976 [143], where it was experimentally shown that when observers were presented with mismatched auditory and visual cues, they perceived a different sound from those presented in the stimulus, i.e. the syllable /ba/ was spoken over the lip movements of /ga/, and the perception was the intermediate syllable /da/. Since then, many authors have demonstrated that the use of visual information in speech recognition improves robustness [183, 184].

Despite audio signals are in general much more informative than video signals, it is known that most people use lip-reading cues to understand speech. However, these cues are often used unconsciously and to different degrees depending on aspects such as the hearing capability [37] or the acoustic conditions (e.g. the visual channel becomes more important in noisy environments) [54], [213], [89], [191]. Furthermore, the visual channel is the only source of information for people with hearing disabilities to understand the oral language if there is no sign language interpreter [200], [183], [10]. In particular, visual information usually involves position and movement of the visible articulators (the lips, the teeth, and the tongue), speaker localization, articulation place, and other signals not directly related to the speech (facial expression, head pose, and body gestures) [89, 236, 37].

In the literature, much of the research has focused on Automatic Speech Recognition (ASR) systems, given that speech is primarily an acoustic form of communication. Nowadays, ASR systems are powerful systems able to understand the spoken language with very high recognition rates when the acoustic signal is not corrupted [39]. However, when the acoustic signal is degraded, the performance of ASR drops and there is the need to rely also on the information provided by the visual channel. This has led to research in

Audio-Visual Speech Recognition (AVSR) systems, which try to balance the contribution of the audio and the visual information channels to develop systems that are robust to audio artifacts and noise. AVSR systems have been shown to significantly improve the recognition performances of audio-based systems under adverse acoustic conditions [183, 51].

On the other hand, in the last decades there has been an increased interest in decoding speech exclusively using visual cues, i.e. mimicking the human capability to perform lip-reading, leading to Visual Speech Recognition (VSR) or Automatic Lip-Reading (ALR) systems [51], [151], [255], [246], [210], [41], [173], [7], [42], [228], [41], [4]. Nonetheless, ALR systems are still behind in performance compared to audio- or audio-visual systems. This can be partially explained by the greater challenges associated to decoding speech through the visual channel, when compared to the audio channel.

One of the main challenges in ALR systems resides on the visual ambiguities that arise at the word level due to homophemes, i.e characters that are easily confused because they produce the same or very similar lip movements (e.g. /p/, /b/ and /m/) [51, 150, 255]. Recall that the main objective of speech recognition systems is to understand verbal communication, which is structured in terms of sentences, words and characters, going from larger to smaller speech entities. More precisely, the standard minimum unit in speech processing is not the character, but the *phoneme*, defined as the minimum distinguishable sound that is able to change the meaning of a word [222]. Similarly, when analyzing visual information many researchers use the *viseme*, which is defined as the minimum distinguishable speech unit in the video domain [66], although there is no consensus on the precise definition of the different visemes nor their number, or even their actual usefulness and existence [32, 66, 44, 194].

The fact that several phonemes produce lip movements that are visually indistinguishable implies that there is no direct or one-to-one correspondence between phonemes and visemes. For example, the phonemes /p/ and /b/ are visually indistinguishable because voicing occurs at the glottis, which is not visible. On the other hand, there are also phonemes whose visual appearance can change (or even disappear) depending on the context: this is the case of the velar consonants (e.g: /k/ or /g/) which change the tongue's position in the palate depending on the previous or following phoneme [146]. For these reasons, many authors have proposed different phoneme-to-viseme mappings, with various definitions and numbers of visemes [18], [87], [152], [63], [99], [27], [7]. In contrast, other authors dispute the existence of visemes and defend that visual ambiguities can be completely resolved using context from neighboring characters, words or a language model [41, 11, 42, 44]. They argue that working through visemes to understand speech is an irrecoverable loss of information. In any case, it is widely accepted that one of the most important challenges when

designing ALR systems is how to make the system robust to visual ambiguities.

Other challenges associated with lip-reading include head pose variations, illumination conditions, poor temporal resolution (when compared to audio systems), efficient encoding of spatio-temporal information and speaker dependency [89, 29, 165]. Furthermore, human lip-readers argue that facial expressions help to decode the spoken message by adding context to the sentence. Thus, while most automatic systems focus only on the mouth region, it might be helpful to consider the whole face to decode visual speech [61].

On the other hand, it is known that some people are very good lip-readers. In general, visual information is the only source of reception and comprehension of oral speech for people with hearing impairments, which leads to the common misconception that they must be good lip-readers. Indeed, while many authors have found evidence that people with hearing impairments outperform normal-hearing people in comprehending visual speech [182, 21, 31, 52, 133], there are also several studies where no differences were found in speech-reading performance between normal-hearing and hearing-impaired people [190, 110]. Such conflicting conclusions might be partially explained by the influence of other factors beyond hearing impairment. For example, it is well known that human lip-readers use the context of the conversation to decode the spoken information [37, 89, 29], thus it has been argued that people who are good lip-readers might be more intelligent, with more knowledge of the language, and with a more comprehensible oral speech for others [145, 190, 165, 111].

Traditionally, ALR systems were based on the extraction of visual features and the classification and modelling of the spoken sequences. Thus, traditional ALR systems mainly consist of image transforms or appearance-based features combined with Hidden Markov Models (HMMs) that use short context information to model the temporal dynamics of the sequences. Early ALR systems addressed simple recognition tasks such as alphabet or digit recognition, but progressively shifted to more complex and realistic settings leading to several recent systems that target continuous lip-reading. To a large extent, these advances have been possible thanks to the construction of powerful systems based on Deep Learning (DL) architectures that have quickly started to replace traditional systems and to the availability of large-scale databases [42, 41].

Despite the recognition rates for continuous lip-reading may appear modest in comparison to those achieved by audio-based systems, the field has undeniably made a significant step forward. Interestingly, an analogous effect can be observed when humans try to decode speech: given sufficiently clean signals, most people can effortlessly decode the audio channel, but would struggle to perform lip-reading, since the ambiguity of the visual cues makes it necessary the use of further context to decode the message. Thus, it is not surprising that the main challenges in ALR systems regard the robustness to visual ambiguities

through the modeling of context information. Most recent works suggest that the optimal modeling of temporal sequences is still an open problem, which is currently been tackled by means of recurrent neural networks or transformers.

1.1 Contribution

In this thesis, we explore the appropriate modeling of visual representations with the aim to properly decode continuous lip-reading. To this end, we present different data-driven mechanisms to handle the main challenges in lip-reading related to the ambiguities of the visual cues and the speaker dependency.

In **Chapter 3**, we explore the fully automatic construction of phoneme-to-viseme mappings based on simple merging rules and the minimization of pair-wise confusion to maximize word recognition. Our experiments support the advantage of merging groups of phonemes into visemes, obtaining the best word accuracy for phoneme-to-viseme mappings with intermediate lengths. Concretely, we highlight that even though going through visemes may seem like a loss of information, this is only partially true because there is no perceivable difference, in visual terms, between some phonemes, and once the viseme classes have been estimated we can recover speech by using higher-level context. While all those complexities may provide some explanation for the rather low recognition rates of ALR systems, there seems to be a significant gap between these and human lip-reading abilities. More importantly, it is not clear what would be the upper bound of visual-speech recognition, especially when the available context is limited (it has been argued that humans can *read* only around 30% of the information from the lips, and the rest is filled-in from the context [165, 50]). Thus, it is not clear if the poor recognition rates of ALR systems are due to inappropriate or incomplete design or because there is an intrinsic limitation in visual information that causes the impossibility of perfect decoding of the spoken message.

In **Chapter 4**, we explore the feasibility of visual speech reading with the aim to estimate the recognition rates achievable by human observers under favourable conditions and compare them with those achieved by an automatic system. To this end, we focus on the design and acquisition of an appropriate database in which recorded speakers actively aim to facilitate lip-reading but conversation context is minimized. Our results suggest that the gap between human lip-reading and automatic speechreading might be more related to the modelling of short and long context than to the ability to interpret mouth appearance.

Therefore, we focused our research on the design of ALR systems that can model different levels of context, i.e. character-, word- and sentence-levels. In this direction, Deep Neural Networks (DNNs) stand out as powerful networks

that properly model spatio-temporal dynamics, though at the expense of requiring massive amounts of training data. Unfortunately, in lip-reading, data has been so far an important limitation, especially for languages different from English. Therefore, in **Chapter 5**, we explore the design of an end-to-end ALR system that can be trained with small-scale data by doing the appropriate restrictions on the learning process of the visual front-end objective. To this end, we introduce a self-supervised training strategy that takes advantage of intermediate labels, named visual units. These visual units are informative enough about the mouth and lips position and are generated in a fully-automatically manner without any human intervention. Our results are competitive with the state-of-the-art and arguably the best to date for this volume of training material.

Finally, in **Chapter 6**, motivated by the fact that every person has unique mouth movements, making the generalization of visual models very difficult, we are the first to explore the unsupervised visual domain adaptation of a Speaker Independent (SI) AVSR system to an unknown and unlabelled speaker. Our assumption is that in an ideal SI system, the same speech events should be represented in a similar way, independently of the speaker. Therefore, we propose to learn a joint distribution between corresponding lip-images in two independent domains. We adapt an AVSR system trained in a source domain to decode samples in a target domain without the need for labels in the target domain. Our results show that the adaptation of the visual front-end to a new speaker benefits the contribution of the visual domain in an SI-AVSR system.

1.2 Significance

Researchers have spent many years studying how machines can mimic several aspects of human behavior, from which speech recognition and synthesis has attracted considerable attention. Speech recognition has primarily been treated as an auditory form of communication, which have been widely investigated for more than fifty years [49], and where nowadays, we can easily find powerful systems already integrated into our societies, e.g. into our computers or smartphones. However, speech is a multi-sensory process that involves both acoustic and visual cues, i.e. most people use lip-reading to some degree as complementary information to understand speech.

Therefore, research on visual-only speech recognition stands out as a novel interesting topic that would be principally beneficial for the development of speech systems in general as well as for people with hearing disabilities. Indeed, there are around 500 million people worldwide who have hearing issues, and the number increases every year. Aspects such as the enormous exposure to noisy environments or the use of headphones for a long time and or at high volume

because tend to increase our possibility to lose, partially or completely, our hearing capability.

1.2.1 Speech Recognition Systems

The use of audio-visual information simultaneously has been proved to provide better results than each modality working separately. This effect is evident because both modalities are complementing one another. Despite AVSR systems try to balance the contribution of the audio and the visual information channels, they are still dominated by the audio modality. Then, the possibility of lip-reading being successful on its own will beyond doubt benefit audio-visual based systems.

Moreover, although audio-based speech recognition is integrated into our daily lives, e.g. our smartphones have a voice recognizer that is able to decode speech to text in real-time, there are a few scenarios where it is impractical or non-sense to depend on a microphone. A real-life example could be an interactive machine in a noisy environment, where there is not a good enough signal-to-noise ratio, e.g. in a subway, an airport, or even in the middle of the street of a busy city. In these situations, the integration of the visual modality would improve speech recognition given that visual speech is not subject to the same ambient noise. Other applications where lip-reading attracts attention are dictating messages to smartphones in noisy environments [72, 209], using visual silent passwords [118, 139, 199, 217, 125], discriminating between native and non-native speakers [75, 74, 76], transcribing and re-dubbing silent films [41, 11], synthesizing voice for people with speech disabilities based on their lip movements [53, 98, 23, 73, 197], developing augmented lip views to assist people with hearing impairments [28], resolving multi-talker simultaneous speech [158, 3], among others.

1.2.2 Hearing-impaired people

Visual speech recognition systems have been especially addressed for helping people with hearing disabilities with the aim to develop systems for the integration and accessibility of these people into society. Cultural activities are still inaccessible in many cases for people with hearing disabilities. For example, the availability of cinemas that offer original version films with subtitles or theater productions accessible with subtitles or sign language interpreters is still very limited.

In this way, access to real-time lip-reading systems combined with intelligent glasses could facilitate the integration of hearing-impaired people to many environments without altering their use by normal-hearing people.

Furthermore, research on lip-reading teaching based on how automatic lip-reading systems learn could be also beneficial to develop systems that train people with hearing disabilities in lip-reading tasks.

1.3 Outline of the thesis

The thesis is organized in 7 chapters. Chapters 2-6 are self-contained and each of them corresponds to a published or under review paper, while Chapter 7 summarizes the conclusions from this work.

Chapter 2. In this chapter, we review ALR research during the last decade, highlighting the progression from approaches previous to DL (which we refer to as traditional) toward end-to-end DL architectures. We provide a comprehensive list of the audio-visual databases available for lip-reading, describing what tasks they can be used for, their popularity, and their most important characteristics, such as the number of speakers, vocabulary size, recording settings, and total duration. In correspondence with the shift toward DL, we show that there is a clear tendency toward large-scale datasets targeting realistic application settings and large numbers of samples per class. On the other hand, we summarize, discuss, and compare the different ALR systems proposed in the last decade, separately considering traditional and DL approaches. We address a quantitative analysis of the different systems by organizing them in terms of the task that they target (e.g. recognition of letters or digits and words or sentences) and comparing their reported performance in the most commonly used datasets. We provide a detailed description of the available ALR systems based on end-to-end DL architectures and identify a tendency to focus on the modeling of temporal context as the key to advance the field.

Chapter 3 In this chapter, we focus on the automatic construction of a phoneme-to-viseme mapping based on visual similarities between phonemes to maximize word recognition. We investigate the usefulness of different phoneme-to-viseme mappings, obtaining the best results for intermediate alphabet lengths. We construct an automatic system that uses DCT and SIFT descriptors to extract the main characteristics of the mouth region and HMMs to model the statistic relations of both viseme and phoneme sequences. We test our system in two Spanish corpora.

Chapter 4. In this chapter, we study the limit of visual speech recognition in controlled conditions. With this goal, we designed a new database in which the

speakers are aware of being read and aim to facilitate lip-reading. In the literature, there are discrepancies on whether hearing-impaired people are better lip-readers than normal-hearing people. Then, we analyze if there are differences between the lip-reading abilities of 9 hearing-impaired and 15 normal-hearing people. Finally, human abilities are compared with the performance of a visual automatic speech recognition system.

Chapter 5. In this chapter, we propose to train an end-to-end ALR system with challenging data by doing the appropriate restrictions on the learning process of the visual front-end objective. To this end, we introduced a self-supervised training strategy that takes advantage of intermediate labels, named visual units. These visual units are informative enough about the mouth and lips position and are generated in a fully-automatically manner without any human intervention. Additionally, we also present a data augmentation strategy that allows synthesizing novel realistic video sequences by appropriately combining characters-like sub-sequences from existing videos.

Chapter 6. In this chapter, we focus on AVSR which faces the difficult task of exploiting acoustic and visual cues simultaneously. Augmenting speech with the visual channel creates its own challenges, e.g. every person has unique mouth movements, making the generalization of visual models very difficult. This factor motivates our focus on the generalization of speaker-independent (SI) AVSR systems especially in noisy environments by exploiting the visual domain. Specifically, we are the first to explore the visual adaptation of an SI-AVSR system to an unknown and unlabelled speaker. We adapt an AVSR system trained in a source domain to decode samples in a target domain without the need for labels in the target domain. For the domain adaptation of the unknown speaker, we use Coupled Generative Adversarial Networks to automatically learn a joint distribution of multi-domain images.

Chapter 7. Finally, in this chapter, we summarize this thesis by giving the most important ideas and contributions of the work.

1.4 Publications from this thesis

Journals

- **Fernandez-Lopez, A., & Sukno, F. M. (2020).** End-to-end Lip-Reading without Large-Scale Data. *International Journal of Computer Vision.* (Under Review)

- **Fernandez-Lopez, A.,** & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, 78, 53-72. DOI: 10.1016/j.imavis.2018.07.002

International Conferences

- **Fernandez-Lopez, A.,** Karaali, A., Harte, N., & Sukno, F. M. (2020, May). Cogans For Unsupervised Visual Speech Adaptation To New Speakers. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6294-6298). IEEE. DOI: 10.1109/ICASSP40776.2020.9053299. (Oral presentation)
- **Fernandez-Lopez, A.,** & Sukno, F. M. (2019, September). Lip-Reading with Limited-Data Network. In *2019 27th European Signal Processing Conference (EUSIPCO)* (pp. 1-5). IEEE. DOI: 10.23919/EUSIPCO.2019.8902572. (Oral presentation)
- **Fernandez-Lopez, A.,** Martinez, O., & Sukno, F. M. (2017, May). Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 208-215). IEEE. DOI: 10.1109/FG.2017.34.
- **Fernandez-Lopez, A.,** & Sukno, F. M. (2017). Automatic phoneme-to-viseme mapping construction to enhance continuous lip-reading. In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)-Volume 5: VISAPP; 2017 Feb 27-Mar 1; Porto, Portugal. Setúbal, Portugal: SCITEPRESS, 2017. p. 52-63. SCITEPRESS. DOI: 10.5220/0006102100520063. (Oral presentation)*

Book Chapters

- **Fernandez-Lopez, A.,** & Sukno, F. M. (2019). Optimizing Phoneme-to-Viseme Mapping for Continuous Lip-Reading in Spanish. In Cláudio, A. P., Bechmann, D., Richard, P., Yamaguchi, T., Linsen, L., Telea, A., Imai, F., and Tremeau, A. (Ed.). *Computer Vision, Imaging and Computer Graphics - Theory and Applications*, (pp. 305-328). Cham. Springer International Publishing. DOI: 10.1007/978-3-030-12209-6_15.

Chapter 2

REVIEW OF THE LITERATURE

In this chapter, we review the research on Automatic Lip-Reading (ALR) systems between 2007 and 2017¹, highlighting the progression from approaches previous to Deep Learning (DL) (which we refer to as traditional) toward end-to-end DL architectures. We provide a comprehensive list of the audio-visual databases available for lip-reading, describing what tasks they can be used for, their popularity and their most important characteristics, such as the number of speakers, vocabulary size, recording settings and total duration. On the other hand, we summarize, discuss and compare the different ALR systems proposed in the last decade, separately considering traditional and DL approaches. We address a quantitative analysis of the different systems by organizing them in terms of the task that they target and comparing their reported performance in the most commonly used datasets.

The remainder of this chapter is organized as follows: in Section 2.1 we summarize the available corpora for lip-reading and their main characteristics, grouped by recognition task and viewing angle. In Section 2.2) we review the progression of ALR systems in the last decade in terms of system architecture and performance, including: i) a review of traditional architectures grouped by task and dataset, and ii) a review of recent ALR systems based on DL architectures. Conclusions are provided in Section 2.3.

Adapted from: Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, 78, 53-72. DOI: 10.1016/j.imavis.2018.07.002

¹We also include the works published so far during 2020.

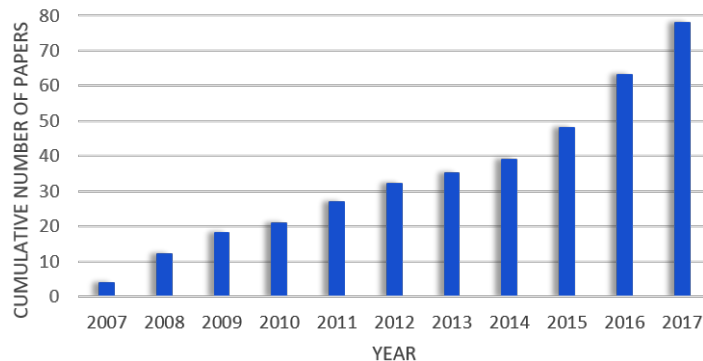


Figure 2.1: Cumulative number of papers on ALR systems published between 2007 and 2017.

2.1 Audio-Visual databases

Reviewing the literature, the early databases designed to develop ALR systems, starting from the nineties, focused on specific and simple recognition tasks with restricted vocabularies, such as the alphabet or digit recognition. These datasets have been widely analyzed because they allow to quickly train prototype systems given that they tackle lip-reading from well-controlled settings with a pre-defined vocabulary and multiple repetitions. However, the typically low numbers of subjects and the limited amount of recorded data make it difficult to construct robust ALR models that generalize well to more realistic application settings. Thus, subsequent databases focused on increasing the amount of captured data and addressing more complex tasks, going toward ALR systems targeting continuous speech. Acquisition of large audio-visual databases is challenging due to the several factors that could be addressed (subjects, repetitions, illumination, head-pose, vocabulary, resolution, etc). Thus, some efforts were made to create datasets providing moderately large amounts of data focusing just on a few factors, while giving up other aspects. For example, the GRID corpus [46] contains a big number of utterances but very similar and constrained sentences and the RM-3000 database [93] contains only one speaker but it has a huge vocabulary. More recent efforts have led to large-scale databases collected from TV broadcasts with the objective to provide a wide vocabulary under increasingly realistic settings (LRW [42], LRS [41], MV-LRS [44]). The biggest dataset for continuous speech recognition, named LRS, consists of more than 100,000 utterances spoken by over a thousand different people. Thus, the field is growing toward large databases with a lot of variability to train robust ALR systems.

In the following subsections, we compare the available databases for training ALR systems, classifying them by task (e.g. letters, digits, words and sentences) and by viewing angle. Despite audio-visual datasets have been dominated by frontal-view recordings, ALR systems should deal with multi-view lip-reading to decode speech in realistic scenarios. Table 2.1 provides a list of audio-visual databases for ALR with frontal-view data, while Table 2.3 provides a similar list for datasets captured under multiple viewpoints. For each database we summarize its key features, including: year of creation; language; the number of speakers; recognition task being considered; the number of classes; the number of utterances; resolution and total duration. In addition, representative snapshots from some of these databases are shown in Figure 2.2.

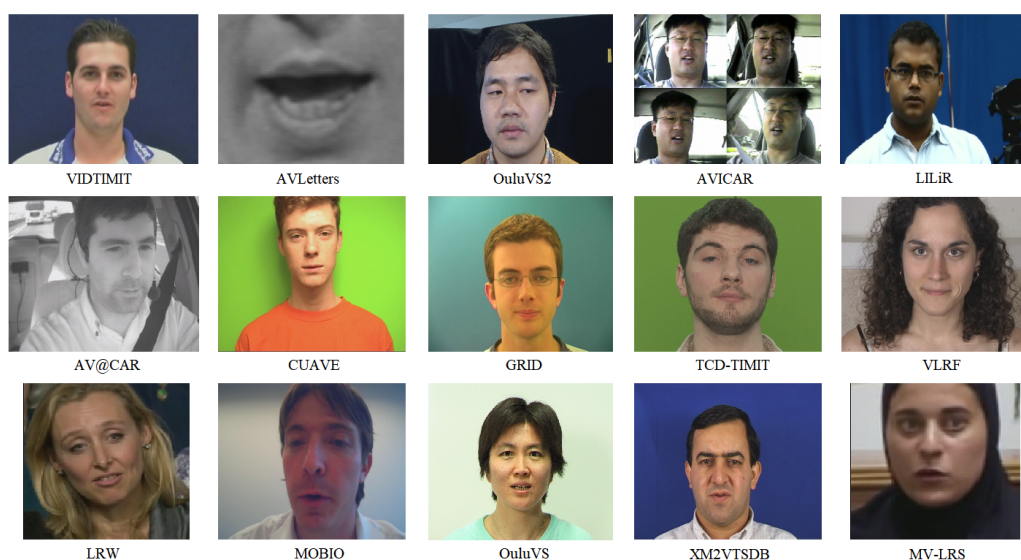


Figure 2.2: Example shots of audio-visual speech databases.

2.1.1 Alphabet and digit recognition

Early works in ALR focused on simple recognition tasks such as alphabet or digit recognition. The available databases differ in several aspects, such as number of speakers, language, number of utterances and spatial and temporal resolutions.

For alphabet recognition, AVLetters (1998) [140] is one of the most used databases. It contains recordings from 10 speakers repeating each letter 3 times, at a resolution of 376×288 pixels and 25 fps. Later on, AVLetters2 [47] and AVICAR [116] solved some weaknesses of AVLetters, such as the low resolution or the limited number of speakers. Specifically, AVLetters2 increased the number of utterances (from 3 to 7 repetitions per speaker) and the resolution (1920×1080

pixels and 50 fps). Nonetheless, the number of speakers was reduced to just 5. On the other hand, AVICAR is a large multi-speaker database with high resolution. It contains 100 speakers, although only 86 are available.

For digit recognition, XM2VTS [144] is one of the biggest multi-speaker databases with 295 participants. It was especially designed for personal identification. Each subject was asked to pronounce two continuous digit strings and one phonetically balanced sentence. Other databases such as VALID [67] or BANCA [13] followed a similar structure to the XM2VTS database. In particular, VALID was designed for comparing speaker identification experiments under controlled and uncontrolled illumination and acoustic noise. This database includes recordings from 106 speakers in five scenarios. Similarly, the BANCA database was especially designed for identity verification under 3 different scenarios (controlled, degraded and adverse). It consists of 208 subjects covering 4 different languages (English, French, Italian and Spanish). There are 12 sessions per subject in which they were instructed to say a random 12 digit number, his/her name, their address and birth-date ($\sim 30,000$ utterances).

However, the most popular database for training ALR systems in digit recognition is CUAVE [170] despite it contains considerably less speakers than XM2VTS and VALID. CUAVE contains 36 speakers but it provides a large number of utterances, organized in sessions of single and dual speakers. In single-speaker sessions, the speaker pronounced 50 isolated digits while standing naturally in front of the camera. After that, the speaker was captured from both profile views while uttering 20 isolated digits, and then 60 connected digits facing the camera again. For dual-speaker sessions, two speakers were recorded at the same time; while one speaker was talking the other one would remain silent, but both were captured by the camera. Speakers were asked to utter two repetitions of connected-digit sequences, alternating their turns. Subsequent datasets were presented dealing with digit recognition such as AV@CAR [164] for Spanish, AVOZES [78], AVICAR [116] and AusTalk [58] for English, the AGH AV Corpus [97] for Polish and the CENSREC-1-AV [216] for Japanese. They were recorded with moderate spatial and temporal resolutions and at least 20 speakers. Other datasets such as IBMIH [96] and IBMSR [128] were designed for digit recognition with huge numbers of speakers and utterances, but unfortunately they are not publicly available. In 2015, the multi-view OuluVS2 database [9] was presented with high resolution, 52 subjects and near 1,600 utterances. More recently, in 2018 the multi-view AV Digit database [174] was presented also with high resolution, 53 subjects and close to 800 utterances of digit sequences.

Table 2.1: Audio-visual corpora, in chronological order

Name	Year	Language	Speakers	Task	Classes	Utterances	Resolution	Duration
AVLetters [140]	1998	English	10	Alphabet	26	780	376×288, 25 fps	13 min
XM2VTS [144]	1999	English	295	Digits	10	885	720×576, 25 fps	59 min
IBMViaVoice [152]	2000	English	290	Sentences	10,500	24,325	704×480, 30 fps	50 h
VIDTIMIT [198]	2002	English	43	Sentences	346	430	512×384, 25 fps	30 min
BANCA [13]	2003	Multiple	208	Digits	10	29,952	720×576, 25 fps	~ 14 h
IBMIH [96]	2004	English	79	Digits	10	16,197	720×480, 30 fps	N/A
AVOZES [78]	2004	English	20	Digits	10	200	720×480, 30 fps	~ 2 h
				Sentences	3	60		
				Alphabet	26	800		
AV@CAR [164]	2004	Spanish	20	Digits	10	600	768×576, 25 fps	~ 1 h
				Sentences	250	6,000		50 min
				Alphabet	26	6,000		~ 8 h
AVICAR [116]	2004	English	86	Digits	13	59,000	720×480, 30 fps	~ 33 h
				Sentences	1317 [†]			
				Digits	10			
CUAVE [170]	2004	English	36	Digits	10	7,000	720×480, 30 fps	14 min
AV-TIMIT [87]	2004	English	233	Sentences	510	4,660	720×480, 30 fps	4 h
VALID [67]	2005	English	106	Digits	10	1,590	576×720, 25 fps	N/A
GRID [46]	2006	English	34	Phrases	51 [†]	34,000	720×576, 25 fps	~ 28 h
IBMSR [128]	2008	English	38	Digits	10	1,661	368×240, 30 fps	N/A
AVLetters2 [47]	2008	English	5	Alphabet	26	910	1920×1080, 50 fps	15 min
IV ² [179]	2008	French	300	Sentences	15	4,500	780×576, 25 fps	~ 8 h
UWB-07-ICAV [221]	2008	Czech	50	Sentences	7,550	10,000	720×576, 50 fps	25 h
OuluVS [250]	2009	English	20	Phrases	10	1,000	720×576, 25 fps	16 min
CENSREC-1-AV [216]	2010	Japanese	42	Digits	10	3,234	720×480, 30 fps	N/A
QuLips [169]	2010	English	2	Digits	10	3,600	720×576, 25 fps	N/A
				Words	6,907	6,907		
NDUTAVSC [38]	2010	German	66	Sentences	6,907	6,907	640×480, 100 fps	~ 11 h
WAPUSK20 [227]	2010	English	20	Phrases	52	2,000	640×480, 32 fps	20 h
LILiR [115]	2010	English	12	Sentences	200	2,400	720×576, 25 fps	N/A
BL [19]	2011	French	17	Sentences	238	4,046	640×480, 30 fps	~ 6 h
UNMC-VIER [237]	2011	English	123	Sentences	12	2,460	708×640, 29 fps	N/A
MOBIO [142]	2012	English	150	Sentences	N/A	N/A	640×480, 16 fps	61 h
AGH AV [97]	2012	Polish	20	Digits	N/A	N/A	1920×1080, 50 fps	~ 3 h
MIRACL-VC [188]	2014	English	15	Words	10	1,500	640×480, 15 fps	N/A
				Phrases	10	1,500		
AusTalk [58]	2014	English	1000	Digits	10	24,000	640×480	~ 3000 h
				Words	966	966,000		
				Sentences	59	59,000		
MODALITY [48]	2015	English	35	Words	182	231	1920×1080, 100 fps	N/A
OuluVS2 [9]	2015	English	52	Digits	10	1,590	1920×1080, 30 fps	~ 1 h
				Phrases	530	530		~ 1 h
				Sentences	530	530		13 min
RM-3000 [93]	2015	English	1	Sentences	1,000 [†]	3,000	360×640, 60 fps	~ 4 h
IBM AVSR [148]	2015	English	262	Sentences	10,400 [†]	N/A	704×480, 30 fps	~ 40 h
TCD-TIMIT [86]	2015	English	62	Sentences	5,954	6,913	1920×1080, 30 fps	~ 6 h
HAVRUS [225]	2016	Russian	20	Sentences	1,530	4,000	640×460, 200 fps	N/A
LRW [42]	2016	English	1,000+	Words	500	400,000	256×256, 25 fps	~ 111 h
LRS [41]	2017	English	1,000+	Sentences	17,428 [†]	118,116	160×160, 25 fps	~ 33 h
VLR [61]	2017	Spanish	24	Sentences	1,374 [†]	10,200 [†]	1280×720, 50 fps	~ 3 h
MV-LRS [44]	2017	English	1,000+	Sentences	14,960	74,564	160×160, 25 fps	~ 20 h
AV Digits [174]	2018	English	53	Digits	10	795	1280×780, 30 fps	N/A
			39	Phrases		5,850		
LRS2 [41]	2017	English	1,000+	Sentences	62,769 [†]	144,482	N/A	438 h
LRS3 [5]	2018	English	9,545	Sentences	70,136 [†]	152,452	N/A	438 h
VoxCeleb2 [40]	2018	6,112	Sentences	N/A	1,128,246	N/A	N/A	2,442 h [*]
YT31k [134]	2019	English	N/A	Sentences	N/A	N/A	N/A	31,000 h
LSVSR [203]	2019	English	N/A	Sentences	127,055	2.9M	N/A	3886 h

[†] Number of words

* Annotations unavailable

h: hours, min: minutes.

Table 2.2: Sentence examples of audio-visual databases

Name	Year	Language	Sentences or Phrases
AVICAR	2004	English	This was easy for us.
			First add milk to the shredded cheese.
			Tofu is made from processed soybeans.
GRID	2006	English	Bin blue at A 1 again.
			Lay green by B 2 now.
			Place red in C 3 please.
OuluVS	2009	English	Excuse me.
			Nice to meet you.
			How are you.
VIDTIMIT	2009	English	She had your dark suit in greasy wash water all year.
			Don't ask me to carry an oily rag like that.
			The clumsy customer spilled some expensive perfume.
UNMC-VIER	2011	English	Joe took father's green shoe bench out.
			She had your dark suit in greasy wash water all year.
			Mum strongly dislikes appetizers.
OuluVS2	2015	English	Military personnel are expected to obey government orders.
			Agricultural products are unevenly distributed.
			Chocolate and roses never fail as a romantic gift.
TCD-TIMIT	2015	English	She had your dark suit in greasy wash water all year.
			The prospect of cutting back spending is an unpleasant one for any governor.
			Don't ask me to carry an oily rag like that.
VLRf	2017	Spanish	Eligieron una casa allí con las mismas condiciones.
			Los gusanos son animales invertebrados sin extremidades.
			A las ocho de la mañana ya estaba haciendo pasteles.
LRS	2017	English	When you're cooking chips at home.
			The traditional chip pan often stays on the shelf.
			Through what they call a knife block.

2.1.2 Word and sentence recognition

Datasets for digit and alphabet recognition have been very popular because they allow dealing with ALR under controlled settings with a constrained vocabulary and large numbers of instances per class. While this is useful to analyze the effectiveness of algorithms at early design stages, the resulting models tend to be of limited scope and difficult to extrapolate to more complex tasks such as word or sentence recognition. However, the aim of ALR systems is to understand natural speech, which is mainly structured in terms of sentences, which has made it necessary for the acquisition of databases containing words, phrases and phonetically-balanced sentences.

One of the earliest audio-visual databases containing sentences is IBMViaVoiceTM [152], which consists of 290 subjects uttering continuous speech read from a script with a vocabulary size of approximately 10,500 words and 24,325 sentence utterances. Unfortunately, this corpus is not publicly available. Among the available corpora we find VIDTIMIT (2002) [198], designed to target person verification. It consists of 43 subjects reciting 10 sentences each, selected from a pool of 346 different sentences. Similarly,

AV-TIMIT [87] was published in 2004 for audio-visual speech recognition. It contains 233 speakers and 510 different sentences. Other datasets already described in Section 2.1.1 for digit recognition also contain specific sessions with sentences: AV@CAR provides 250 phonetically-balanced sentences, AVICAR sentences with more than 1,300 different words, and AVOZES three different sentences designed to contain almost all phonemes and visemes of Australian English.

Several other databases were published between 2008 and 2014. Most of them were recorded in English [114], [250], [237], [142], [188], [58] but we can also find two databases recorded in French [179] and one recorded in Czech [221]. Among the English-based corpora, the OuluVS database [250] is one of the most used databases for evaluating ALR systems. It contains 20 speakers uttering 10 short sentences of daily-use in English, where each utterance was repeated by the same speaker up to 5 times. The LILiR [115], MIRACL-VC [188], UNMC-VIER [237] and Austalk [58] databases contain 12, 15, 123 and 1000 speakers, respectively. However, MIRACL-VC and UNMC-VIER contain rather few sentences (10 and 12), while LILiR and Austalk contain 200 and 59 different sentences, respectively. Yet within English corpora, we also find the MOBIO database [142]. Differently from those previously mentioned, the MOBIO database was designed for evaluating automatic face and speaker recognition on a mobile phone. It contains videos from 150 speakers answering short and free-speech questions and reading predefined texts, always recorded with a mobile phone held by themselves.

Audio-visual databases recorded in other languages are much less frequent than those in English. For example, in the French language we find the IV² [179] and BL [19] databases; the first one provides a large number of speakers (300) uttering 15 sentences, while BL provides just 17 speakers but 238 sentences each. Other examples include the UWB-07-ICAVR database [221], which provides 10,000 utterances from 50 subjects in Czech, the NDUTAVSC database [38], with 66 German speakers, the AV@CAR database [164], in Spanish (already described above) and the VLRf database [61], also in Spanish, providing 1,507 utterances from 24 speakers. In Table 2.2 we show examples of sentences of some of these AV-databases.

More recently, other databases have been published. Among them we find the single speaker RM-3000 corpus [93] which contains a vocabulary of 1,000 different words and 3,000 utterances. In contrast, we find several multi-speaker databases, namely OuluVS2 [9], TCD-TIMIT [86], HAVRUS [225], IBM AVSR [148], VLRf [61] and AV Digits [174], which contain 53, 62, 20, 262, 24 and 53 subjects, respectively. OuluVS2 contains recordings of speakers uttering phrases and sentences; each speaker repeated three times a set of 10 daily-use phrases (similar to OuluVS) and read 10 TIMIT sentences randomly chosen from a total

of 530 sentences. On the other hand, the TCD-TIMIT dataset contains more than 6,900 sentences while the HAVRUYS database [225], in Russian, provides 4,000 utterances from 20 speakers. The IBM AVSR database is a large corpus whose sentences contain more than 10,000 words, but unfortunately it is not publicly available. The VLRf database, in Spanish, contains 24 speakers repeating up to three times sets of 25 sentences selected from a pool of 500 phonetically-balanced sentences (10,000+ word utterances). Interestingly, this corpus includes participants with different hearing capabilities: 15 were normal-hearing and 9 were hearing-impaired subjects, who also performed lip-reading on the recorded videos. The transcriptions of the human lip-reading are also provided, allowing for a direct comparison between human and ALR. Finally, the very recent AV Digits database contains videos of 39 speakers uttering 10 daily-use phrases (similar to OuluVS and OuluVS2). Each phrase is repeated five times in three different speech modes: normal, whispered and silent.

Another key element to consider is the widespread use of Deep Neural Networks (DNNs) in the last few years, which has produced important advances in many aspects of computer vision, including of course lip-reading systems. While these networks have demonstrated considerable improvements in classification performance, this is only possible if appropriate data are available for training. In other words, DNNs are characterized by the need for big amounts of training data. Even though we have mentioned numerous audio-visual databases suitable for ALR, most of them do not contain a sufficient number of samples or do not cover enough vocabulary to train DNNs that generalize well. Thus, early attempts of ALR systems based on DL faced a shortage of data and, among the available corpora, those with a larger number of utterances per class became more popular. For example, the GRID corpus [46] was introduced in 2006 but its use has considerably increased in the last few years. This corpus contains data collected from 34 speakers uttering 1,000 constrained sentences, each fitting into a 3-second time window. Each speaker produced all combinations of "color", "digit" and "letter" by following the fixed sentence structure <command> + <color> + <preposition> + <digit> + <letter> + <adverb>. It contains 34,000 utterances of very similar sentences with a vocabulary that covers 51 words. There exist also other databases that follow a similar sentence structure such as WAPUSK20 [227] or MODALITY [48]. These corpora provide rather large number of instances per class, which is adequate for training DNNs, but cannot generalize outside of the rather small set of words that they cover.

Therefore, new databases have been recently recorded with the aim of providing both large numbers of utterances and a wider vocabulary. Among these, most relevant efforts include the LRW [42], LRS [41] and MV-LRS [44]

Table 2.3: Multi-view audio-visual databases, in chronological order

Name	Year	Language	Task	Speakers	Classes	Utterances	View (°)
CUAVE [170]	2004	English	Digits	36	10	7,000	-90, 0, 90
AVICAR [116]	2004	English	Sentences	100	1,317 [†]	59,000	Variable (4 views)
CMU AVPFV [108]	2007	English	Words	10	150	15,000	0, 90
IBMSR [128]	2008	English	Digits	38	10	1,661	-90, 0, 90
HIT-AVDB-II [245]	2008	Multiple(2)	Sentences	30	11	1,980	0, 30, 60, 90
QuLips [169]	2010	English	Digits	2	10	3,600	0, 10, 20, ..., 90
LILiR [115]	2010	English	Sentences	12	200	2,400	0, 30, 45, 60, 90
LTS5 [56]	2011	French	Digits	20	10	180	0, 30, 60, 90
OuluVS2 [9]	2015	English	Sentences	53	540	2,120	0, 30, 45, 60, 90
TCD-TIMIT [86]	2015	English	Sentences	62	6,913	13,826	0, 30
MV-LRS [44]	2017	English	Sentences	3,783	14,960 [†]	74,564	from 0 to 90
AV Digits [174]	2018	English	Digits	53	10	795	0, 45, 90
			Phrases	39		5,850	

[†] Number of words

databases. The Lip Reading Words (LRW) and Lip Reading Sentences (LRS) databases are based on recordings from BBC programs between 2010 and 2016. LRW contains sentences from more than 1,000 speakers and a vocabulary of 500 words that occur at least 800 times each ($\sim 400,000$ utterances in total). LRS contains 17,428 different words combined in 118,116 utterances along with the corresponding facetrack. Finally, the MultiView-LRS (MV-LRS) database was also recorded from BBC programs but, while LRW and LRS contain only frontal face shots, MV-LRS includes shots from any viewing angle between 0 and 90 degrees. Unfortunately, LRS is not publicly available.

2.1.3 Multiview databases

ALR systems have been usually based on visual speech understanding from frontal view recordings. However, in a practical system it is not always possible to ensure that the input images will be exclusively from frontal shots. For example, in the case of imaging multiple speakers in a conversation with a single camera, we will need to work with images from different angles for each speaker. Thus, practical ALR systems should tackle multi-view lip-reading to be able to understand speech in realistic application scenarios. Furthermore, studies with human lip-readers have found that perfectly frontal shots are not necessarily the best ones to perform lip-reading. Indeed, angles slightly departing from frontal-view have shown to be beneficial because lip protrusion and rounding can be better observed [114]. Then, in this section we review datasets that provide speaker recordings from different viewpoints (Table 2.3).

There is considerable variability in the recording setups that have been used to capture multi-view databases for audio-visual research. Some of them contain only frontal and full-profile views, while others contain several slots between 0

and 90 degrees. On the other hand, there are datasets that have been recorded by multiple cameras simultaneously capturing the speaker at different angles, while others have used a single camera to record different views of the speaker sequentially, at different time instants.

The AVICAR database, described in Section 2.1.1, was recorded in a moving automobile using an array of four cameras and eight microphones. The cameras were placed on the dashboard of the car and recorded simultaneously 4 near-frontal views of the driver. Other databases contain recordings from frontal and profile views such as the CUAVE, the CMU AVPFV [108] and the IBMSR databases. CUAVE contains single-camera recordings from people uttering sequences of digits in frontal views and in both profiles (further details in Section 2.1.1). In contrast, the CMU AVPFV database [108] consists of simultaneously-recorded profile and frontal views. It contains data from 10 subjects, with each subject repeating 150 possible words 10 times. Similarly, the IBMSR database, consists of recordings of three cameras simultaneously capturing frontal and two side views from 38 subjects while uttering digits sequences, but unfortunately it is not publicly available.

More recently, several databases have been presented with views between 0 and 90 degrees. For digit recognition, we find the QuLips database [169] and the LTS5 database [56]. QuLips contains recordings from two cameras capturing each speaker while uttering sequences of digits in English (2 speakers in total). The first camera was always kept at the initial position while the subject and the second camera were allowed to rotate, so that different angles at 10° steps could be captured two at a time. In contrast, LTS5 consists of recordings of 20 native French speakers uttering digit sequences. The recordings involve one frontal camera plus one camera rotating to 30° , 60° and 90° relative to the speaker in order to obtain two simultaneous views of each sequence. For each possible position of the second camera, the speaker repeated three times the same digit sequence.

Several multi-view databases have been presented for sentence recognition in English: LILiR [115], OuluVS2 [9], TCD-TIMIT [86], MV-LRS [44], AV Digits [174] and HIT-AVDB-II [245]. Most of them have been recorded by multiple cameras, so that the different views are synchronized. For instance, LILiR contains recordings of 5 cameras located at 0° , 30° , 45° , 60° and 90° while OuluVS2 contains recordings from the same positions as LILiR but using 2 cameras with different resolution for frontal views. Similarly, TCD-TIMIT and HIT-AVDB-II contain recordings with two cameras, one fixed at the frontal view and the other one fixed at 30° for TCD-TIMIT or rotating at 30° , 60° and 90° for HIT-AVDB-II. Interestingly, HIT-AVDB-II provides various types of utterances in English and Chinese. AV Digits contains high-resolution recordings with three cameras, one fixed at the frontal view, another one fixed at 45° and the last one

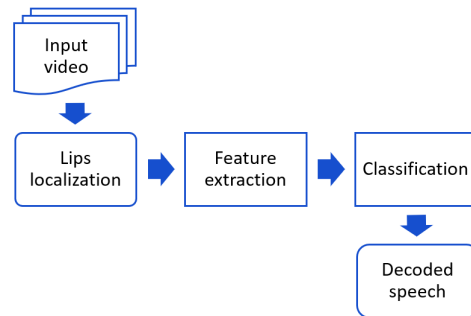


Figure 2.3: The main processing blocks of an ALR system

fixed at the full-profile view. Finally, MV-LRS is based on a selection from a wide range of BBC programs where people engage in conversations with one another, and are therefore more likely to be captured from lateral views. Thus, it contains recordings of people captured at variable views from 0 to 90 degrees; although this dataset does not provide the viewing angle between the speaker and the camera.

2.2 Automatic lip-reading systems

In this section we review the research on ALR systems published between 2007 and 2017. Figure 2.1 provides a quick view of the growth of the field in this period of time, by showing the cumulative number of papers that were published per year. We can observe a significant increase in the number of papers published in the last few years that, as we shall see, coincides with the growing development of DL architectures and the availability of large-scale databases.

Tables 2.4, 2.5 and 2.6 summarize the main characteristics of the ALR systems considered in Figure 2.1. Specifically, we show the publication year, the proposed architecture (in terms of features and classifiers), the database used, the recognition task that was targeted and the accuracy that was reported. Whenever possible, we provide the accuracy in terms of Word Recognition Rates (WRR); otherwise we provide other metrics indicative of ALR performance as provided in the corresponding publications (e.g. phoneme or viseme accuracy and correctness).

An interesting aspect that emerges from the above tables is the shift of ALR systems toward architectures based on DL, which is especially noticeable in 2016 and 2017. Thus, we analyze in separate subsections the approaches previous to DL (which we refer to as traditional) and those that employ DL architectures. In all cases, we focus on the aspects specific to lip-reading and

skip other pre-processing stages more related to face analysis applications in general. Specifically, in Figure 2.3 we show the schematic diagram of a typical ALR system, which consists of three main blocks: 1) Lips localization, 2) Extraction of visual features, 3) Classification into sequences. The first block, focused on face detection and lips localization, will not be covered in this survey; the interested reader is referred to works on face localization and landmarking [226, 248, 138, 162, 163, 12, 193, 247, 223, 239]. The goal of the feature-extraction block is to parametrize the visual information observable at a given time instant or window and the classification block aims to map the visual features into speech units while incorporating temporal constraints to ensure that the decoded message is coherent. The latter provides robustness against noisy or imperfect estimates from the visual cues and helps to disambiguate between visually similar speech units. The rest of the section will focus on the last two blocks: feature extraction and classification.

We review traditional ALR systems in Section 2.2.1 and DL systems in Section 2.2.2. In both cases, we address a quantitative analysis of the different systems by organizing them in terms of the task that they target (e.g. recognition of letters or digits and words or sentences) and comparing their reported performance in the most commonly used datasets. This is important for a fair comparison, given that results are usually reported in different databases, for different recognition tasks, with a variable number of speakers, vocabularies, language and so on. Furthermore, we discuss the most popular DNN architecture for ALR systems and compare several variations that follow this baseline structure. In addition, we comment other DNNs used for lip-reading that explore alternatives from the baseline architecture and provide figures with block diagrams of the most representative end-to-end ALR systems up to 2017.

2.2.1 Traditional ALR systems

ALR systems start by detecting the face and extracting the region that comprises the mouth and its surrounding area. Leaving aside this pre-processing step, once the speaker's lips are located, feature extraction techniques are applied. However, for visual speech recognition, there is no consensus on which is the best feature extraction technique and there are discrepancies, for example, on whether there is more information in the position of the lips or in their movement [130], [194], [32]. Thus, many researchers have proposed ALR systems with different visual features based on image transforms (e.g. DCT), motion (e.g. Optical flow), geometry (e.g. width and height of the mouth) or statistical models (e.g. AAM) [80, 59, 137, 150, 91, 126, 255, 26]. In contrast, most traditional ALR systems use HMMs to classify the visual features into speech units because they help to disambiguate between visually similar speech units while they give linguistic

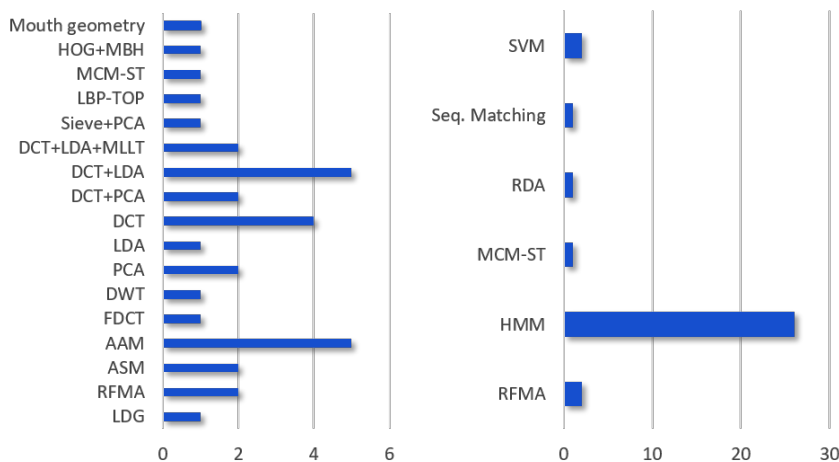


Figure 2.4: Digit and alphabet recognition. Left-side: number of times that each feature technique has been used from 2007 to 2017; Right-side: number of times that each classification method has been used from 2007 to 2017.

consistency to the output message.

Digit and letter recognition

There are 23 ALR architectures targeting digit or alphabet recognition since 2007. Looking at Tables 2.4, 2.5 and 2.6 we observe that most traditional systems use feature techniques based on image transforms [127, 200, 128, 84, 95] or shape and appearance models [47, 168, 167, 89, 232]. In Figure 2.4 we show i) the number of times that each feature technique has been integrated into ALR systems addressing digit or letter recognition; ii) the same for each classification method. On the left-side of the figure, we observe that the most used visual features have been AAMs, DCT or combinations of DCT with other transforms such as LDA or PCA. On the other hand, on the right-side of the figure, a single HMM for each digit or letter is the most used classification method, being also the most used in audio speech recognition. Other methods such as Support Vector Machines (SVM) or Regularized Discriminant Analysis (RDA) have less been frequently explored.

Given the variety of methods addressing digit or letter recognition, it is interesting to compare them in terms of performance. This can be directly done by comparing the methods evaluated in the same databases. Thus, we will compare the methods evaluated in the most commonly used databases for digit or alphabet recognition, which are CUAVE, XM2VTS or AVLetters2.

Architectures presented in [129, 167, 166, 84, 168, 189, 55] have been evaluated using the CUAVE database. These methods reported WRR between

53.12% and 83.00%. For the 5 architectures using HMMs as classification method, two of them used DCT [129] and LDA [84] features, reporting 53.12% WRR and 60.00% WRR, respectively. Similarly, the system presented by Estellers et al. [55] used DCT features and obtained 60.40% WRR. In contrast, both architectures presented by Papandreou et al. [167, 168] used AAM models and reported 75.70% WRR and 83.00% WRR, respectively. The latter is the best WRR reported in this database. Nevertheless, the ALR system proposed by Pachoud et al. [166] based on probabilistic sequence matching classification of macro-cuboids using spatio-temporal SIFT descriptors and local displacements (named MCM-ST features) reported a similar performance (80% WRR). Finally, there is an ALR system presented in 2016 by Rezik et al. [189] that used a combination of Histogram of Oriented Gradients (HOG) and Motion Boundary Histograms (MBH) features and SVM classifiers reporting a performance of 70.10% WRR.

For the XM2VTS database, Seymour et al. [200] presented experiments comparing different image transforms (DCT, PCA, LDA, and FDCT) combined with HMMs and obtained WRR between 85.36% and 87.89%. On the other hand, the ALR system presented by Stewart et al. [209] presented a conventional system based on DCT features and HMMs, reporting 70.00% WRR. The best-performing architecture for XM2VTS used DCT features and HMMs classifiers and reported 87.89% WRR [200].

Finally, for alphabet recognition, AVLetters2 has been one of the most used databases. Several traditional architectures have been proposed with WRR up to 91,80% [47, 89, 171]. For the HMM-based systems, feature extraction techniques such as Sieve filters combined with PCA [47] and AAM [47, 89] have been used. However, the best WRR was reported by the system presented by Pet et al. [171] that consists of an end-to-end system based on Random Forest Manifold Alignment (RFMA), which obtained 91,80% WRR followed by the 75,24% WRR obtained by Hilder et al. [89].

Therefore, even though DCT has been the most implemented feature in ALR systems tackling digit or alphabet recognition, AAM features in combination with HMMs have produced the highest reported WRR.

Word and sentence recognition

Digit and letter recognition has been very popular, but the resulting models cannot be extrapolated to more complex tasks such as word or sentence recognition and hence are of limited applicability. In Figure 2.5 we show the number of ALR architectures targeting *digit or alphabet* and *word or sentence* recognition from 2007 to 2017. In the figure, we can observe a clear tendency from early systems trying to solve easier recognition tasks in controlled

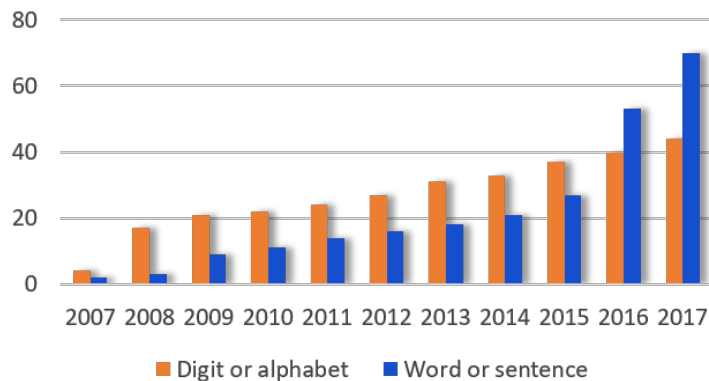


Figure 2.5: Cumulative number of ALR systems targeting *digit or alphabet* and *word or sentence* recognition from 2007 to 2017.

vocabularies (e.g. digits) toward systems dealing with more complex tasks such as word or sentence recognition. In this section we compare the 33 traditional systems presented in Tables 2.4, 2.5 and 2.6 that target word or sentence lip-reading. Similarly to Section 2.2.1, we firstly explain the architecture’s components and then compare systems in terms of performance.

In Figures 2.6 and 2.7 we show, respectively, the number of times that each feature or classification technique has been integrated into ALR systems targeting word or sentence recognition. In Figure 2.6 we observe that the most used visual features are similar to those used in digit or alphabet recognition, namely PCA, DCT, and AAM. Notice that even though these features do not have the highest usage frequencies by themselves, they appear multiple times combined with others. Compared to digit or letter recognition there is a bigger pull of features, e.g. Local Binary Patterns extracted from Three Orthogonal Planes (LBP-TOP), Shape Difference Feature (SDF) or Spatio-Temporal Lip Feature (STLF). In terms of classifiers (Figure 2.7), we also observe a similar tendency to digit or letter recognition, where HMMs are the most used classification method, although there is also an increment of systems using alternative classifiers, especially SVMs.

In terms of performance evaluation, the most used databases for word or sentence recognition have been GRID, OuluVS, OuluVS2 and RM-3000.

For the GRID corpus, Lan et al. [112] used a subset of 15 speakers and centered their experiments in comparing different features such as DCT, Sieve, PCA and AAM. They used one HMM per word for decoding the message, 52 HMMs in total (51 words plus silence). They obtained WRR between 40.00% and 65.00%, being AAM the most successful feature. In contrast, Kolossa et al. [104] proposed a similar model composed of DCT features and one HMM per

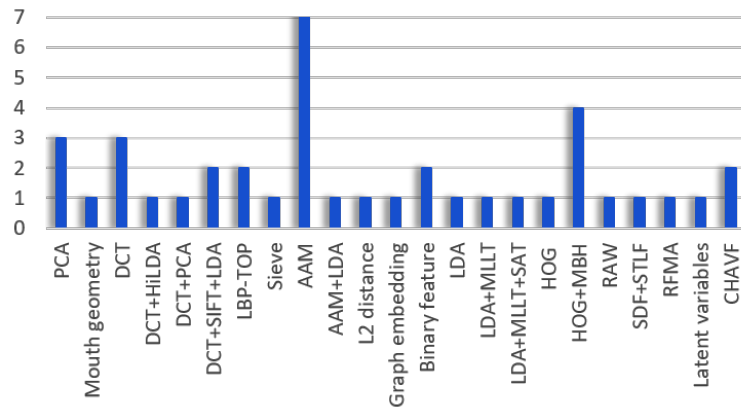


Figure 2.6: Word and sentence recognition. Number of times that each feature technique has been used from 2007 to 2017.

word and reported 57.00% WRR in experiments using the full set of speakers. More recently, Wand et al. [228] compared PCA and HOG using SVM as classifier. They obtained WRR of 69.50% for PCA features and 71.20% for HOG on speaker-dependent experiments over a subset of 20 subjects. Speaker dependent experiments mean that training and testing data for the classifiers are always taken from the same speaker and the results are averaged over all the speakers.

For the OuluVS database, 9 different architectures have been presented [250, 256, 257, 160, 159, 254, 189, 211, 171]. For the ALR systems evaluated in this database, a varied set of features has been used, but most works used SVMs as classifiers. Rekik et al. [189] used a combination of spatio-temporal HOG and MBH features with SVMs and obtained WRR of 68.30%. Sui et al. [211] presented a feature extraction technique named Cascade Hybrid Appearance Visual Feature (CHAVF), which is based on LBP-TOP and DCT features and combined them with SVMs, achieving WRR of 68.90% for speaker-dependent experiments. In contrast, both Zhao et al. [250] and Zhou et al. [257] used LBP-TOP features combined with SVMs and reported 62.40% and 81.30% WRR, respectively. These big difference ($\sim 20\%$) are because Zhou et al. [257] introduced a process of curve matching that normalizes the video signal by mapping the original video onto a curve which is then re-sampled to produce video sequences with the same number of frames. In contrast, Ong et al. [160], [159] proposed two systems based on binary features combined with Temporal Gradient Descend Boosting (TGD-Boosting) [160] or with Sequential Pattern Boosting (SP-Boosting) classifiers [159], reporting 65.60% and 86.20% WRR, respectively. Pei et al. [171] presented an end-to-end system based on RFMA

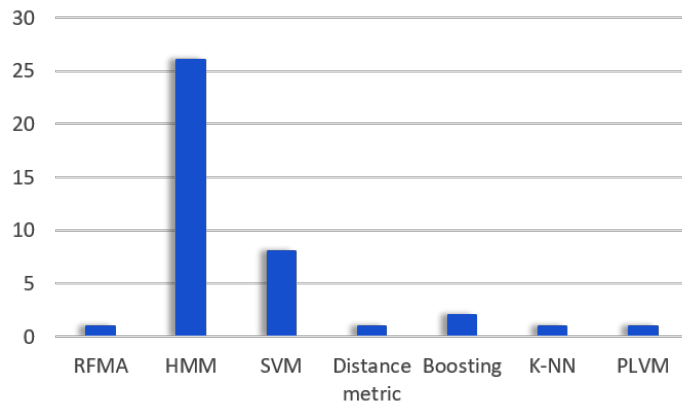


Figure 2.7: Word and sentence recognition. Number of times that each classification method has been used from 2007 to 2017.

and reported 89.70% WRR, which is the highest performance achieved so far in this database. Other alternative systems were presented by Zhou et al. [256, 254]. The first one [256] uses graph embedding to capture video dynamics and the second one [254] used latent variable (LV) models to generate the representation of a sequence of images. For leave-one-utterance-out cross validation in [256] they obtained 90.60% WRR, while for leave-one-speaker-out cross-validation in [254] they obtained 74.00% WRR.

For the OuluVS2 database, Wu et al. [238] presented a feature extraction technique based on SDF and STLF features and SVM classifiers to decode the spoken message, obtaining 55.00% WRR. In contrast, Lee et al. [117] presented three different systems. HMM-based systems were based on DCT-PCA and DCT-HiLDA features and reported 63.00% and 74.00% WRR, respectively, while the third system was based on LV models combined with raw pixel values as features and reported 73.00% WRR.

For the single-speaker RM-3000 dataset with 1000 different words, Thangthai et al. [220] and Howell et al. [92] proposed similar ALR systems using AAM features and HMM classifiers. Thangthai et al. [220] trained Context-Independent HMMs (CI-HMM) and Context-Dependent HMMs (CD-HMM). Instead of directly constructing word models they defined phoneme models. Then, they joined the corresponding phonemes of each word to form word models (model of models). The CI-HMM consisted of monophone models with 3 states per phoneme (45 phonemes in English), while the CD-HMM models distinguished between phonemes with different previous and posterior phonemes. They obtained 33.32% WRR for CI-HMMs and 47.48% WRR for CD-HMMs. Similarly, Howell et al. [92] presented an ALR system based on

AAMs and triphoneme word decoders, and reported a WRR of 75.58%. As we can observe, for databases covering large vocabularies it seems useful to train phoneme or triphoneme models instead of just training words, because this increases the number of samples per class available for training.

For the LILiR database, Bowden et al. [26] proposed a system based on the combination of AAM features and HMM classifiers and obtained 30.20% WRR for one-speaker experiments. Lan et al. [113] used Fisher phoneme-to-viseme mapping [66] and proposed an ALR system that combines AAM+LDA features with HMMs trained on viseme classes, obtaining 14.08% WRR. Almajai et al. [7] also used Fisher phoneme-to-viseme mapping and proposed several CI-HMM and CD-HMM systems. Specifically, they proposed a CI-HMM based on monophone and monoviseme models using first- and second-order derivative features and CD-HMMs based on triphone and triviseme models with LDA, LDA+MLLT and LDA+MLLT+SAT features. In their experiments, they found that when phoneme models are used instead of viseme models, the WRR increases significantly, up to 8%, reaching up to 43.00% WRR for the whole database. Interestingly, the opposite result was reported in [63] for the Spanish database AV@CAR, where a phoneme-to-viseme mapping with an appropriate vocabulary length provided the highest WRR. Thus, there is not a general consensus on whether using visemes is advantageous or disadvantageous for ALR.

Summarizing the systems targeting word or sentence recognition, we have seen that different architectures have been evaluated for each database, both in terms of features and classifiers. In contrast to the case of digit and letter recognition systems, the disparity of features evaluated in each database makes it difficult to conclude which might be the best performing ones. Something similar occurs in terms of classifiers: HMMs reported the best performance for the GRID database, SVMs for the OuluVS database and LV models for the OuluVS2 database. However, no system based on HMMs or LV models was tested in the OuluVS dataset and, although some HMM systems were used for OuluVS2, their features did not match those from the best-performing system. Thus, it is difficult to produce a fair comparison beyond the frequency with which the different features and classifiers have been used.

2.2.2 DNN-based ALR systems

While there is an extensive literature dedicated to hand-crafted methods (Section 2.2.1), there has been a significant improvement in the performance of ALR systems in the last years thanks to the advances in deep neural networks and the availability of large-scale databases.

There is strong parallelism in the way that DNNs have been adopted by

Table 2.4: ALR systems from 2007 to 2017 - Part I

Year	Reference	Model		Database	Recognition task	WRR (%)
		Features	Classifier			
2007	Fu et al. [70]	LDG	HMM	AVICAR	Digits	37.87%
2007	Kumar et al. [108]	Mouth geometry	HMM	CMU AVPFV	Words	32.39% [†]
2007	Lucey et al. [127]	DCT+LDA	HMM	IBMSR	Digits	68.58%
2007	Marcheret et al. [135]	DCT+LDA+MLLT	HMM	IBMIH	Digits	63.00%
2008	Cox et al. [47]	Sieve+PCA	HMM	AVLetters2	Alphabet	83.00%
		AAM	HMM	AVLetters2	Alphabet	85.00%
2008	Lucey et al. [129]	DCT+LDA	HMM	CUAVE	Digits	53.12%
2008	Lucey et al. [128]	DCT+PCA	HMM	IBMSR	Digits	66.21%
2008	Pachoud et al. [166]	MCM-ST	Prob. seq. matching	CUAVE	Digits	80.00%
2008	Papandreou et al. [167]	AAM	HMM	CUAVE	Digits	75.70%
2008	Seymour et al. [200]	DCT	HMM	XM2VTS	Digits	87.89%
		PCA	HMM	XM2VTS	Digits	86.57%
		FDCT	HMM	XM2VTS	Digits	85.36%
		LDA	HMM	XM2VTS	Digits	86.35%
2008	Shao et al. [201]	DCT	HMM	GRID	Phrases	58.40%
2008	Wang et al. [232]	ASM	RDA	Own data	Digits	88.32%
		ASM	HMM	Own data	Digits	91.27%
2009	Gurban et al. [84]	DCT+LDA	HMM	CUAVE	Digits	60.00%
2009	Hilder et al. [89]	AAM	HMM	AVLetters2	Alphabet	75.24%
2009	Kolossa et al. [104]	DCT	HMM	GRID	Phrases	57.00%
2009	Lan et al. [112]	Sieve	HMM	GRID	Phrases	40.00%
		DCT	HMM	GRID	Phrases	40.00%
		Eigenlips	HMM	GRID	Phrases	52.00%
		AAM	HMM	GRID	Phrases	65.00%
2009	Papandreou et al. [168]	AAM	HMM	CUAVE	Digits	83.00%
2009	Zhao et al. [250]	LBP-TOP	SVM	AVLetters	Alphabet	62.80%
		LBP-TOP	SVM	OuluVS	Phrases	62.40%
2010	Pass et al. [169]	DCT	HMM	QuLips	Digits	98.00%
2010	Saitoh et al. [195]	L2 between keypoints	HMM	Own data	Words	68.93%
2010	Zhou et al. [256]	Graph embedding		OuluVS	Phrases	90.60% [†]
2011	Cappelletta et al. [32]	Optical flow	HMM	VIDTIMIT	Sentences	57.00% ^V
		PCA	HMM	VIDTIMIT	Sentences	60.10% ^V
2011	Navarathna et al. [149]	DCT+PCA	HMM	AVICAR	Digits	25.00%
2011	Ngiam et al. [154]	ST-PCA	Autoencoder	AVLetters	Alphabet	64.40%
2011	Ong et al. [160]	Binary feature	TGD-Boosting	OuluVS	Phrases	65.60%
2011	Ong et al. [159]	Binary feature	SP-Boosting	OuluVS	Phrases	86.20%
2011	Zhou et al. [257]	LBP-TOP	SVM	OuluVS	Phrases	81.30%
2012	Chişu et al. [37]	Mouth geometry	HMM	NDUTAVSC	Digits	84.24%
2012	Estellers et al. [55]	DCT	HMM	CUAVE	Digits	60.40%
2012	Estellers et al. [57]	DCT+LDA	HMM	Own data	Digits	71.00%
2012	Lan et al. [114]	AAM	HMM	LILiR	Sentences	33.00% ^V
2012	Lan et al. [113]	AAM+LDA	HMM	LILiR	Sentences	14.08%
2013	Bowden et al. [26]	AAM	HMM	LILiR	Sentences	30.20% [†]
2013	Huang et al. [95]	DCT+LDA	HMM	Own data	Digits	35.20%
		DCT+LDA	DBN	Own data	Digits	35.70%
2013	Pei et al. [171]	RFMA		AVLetters	Alphabet	69.60%
		RFMA		AVLetters2	Alphabet	91.80%
		RFMA		OuluVS	Phrases	89.70%
2014	Bear et al. [18]	AAM	HMM	AVLetters	Alphabet	35.00% ^{C †}
2014	Noda et al. [157]	CNN	MS-HMM	ATR	Words	37.00%
2014	Stewart et al. [209]	DCT	MS-HMM	XM2VTS	Digits	70.00%
2014	Zhou et al. [254]	Latent variables	Cross correlation	OuluVS	Phrases	74.00%
2015	Bear et al. [14]	AAM	HMM	AVLetters2	Alphabet	38.00% ^{C †}
2015	Bear et al. [17]	AAM	HMM	LILiR	Sentences	61.80% ^{C †}
2015	Biswas et al. [22]	AAM	HMM	AVICAR	Sentences	28.23%

* V: Viseme accuracy, P: Phoneme accuracy, C: Correctness.

† Speaker dependent.

Table 2.5: ALR systems from 2007 to 2017 - Part II

Year	Reference	Model		Database	Recognition task	WRR (%)
		Features	Classifier			
2015	Moon et al. [147]	DBN		AVLetters	Alphabet	55.30%
2015	Mroueh et al. [148]	Scattering coeffs+LDA	Feed-Forward	IBM AVSR	Sentences	30.64% ^P
2015	Ninomiya et al. [156]	DBN	MS-HMM	CENSREC-1-AV	Digits	39.30%
2015	Noda et al. [158]	CNN	MS-HMM	ATR	Words	22.50%
2015	Sui et al. [210]	DBM+DCT+LDA	HMM	AusTalk	Digits	69.10%
2015	Thangthai et al. [220]	AAM	CI-HMM	RM-3000	Sentences	33.32%
		AAM	CD-HMM	RM-3000	Sentences	47.48%
		AAM	Feed-Forward	RM-3000	Sentences	77.49%
		HiLDA	Feed-Forward	RM-3000	Sentences	84.67%
2016	Almajai et al. [7]	LDA	HMM	LILiR	Sentences	23.00%
		LDA+MLLT	HMM	LILiR	Sentences	25.00%
		LDA+MLLT+SAT	HMM	LILiR	Sentences	43.00%
		LDA+MLLT+SAT	Feed-Forward	LILiR	Phrases	53.00%
2016	Assael et al. [11]	3D-CNN	Bi-GRU	GRID	Phrases	93.40%
2016	Bear et al. [15]	AAM	HMM-bigram net	LILiR	Sentences	23.00% ^C
2016	Chung et al. [43]	VGG-M	LSTM	OuluVS2	Phrases	31.90%
		SyncNet	LSTM	OuluVS2	Phrases	94.10%
2016	Chung et al. [42]	CNN		LRW	Words	61.10%
		CNN		OuluVS	Phrases	91.40%
		CNN		OuluVS2	Phrases	93.20%
2016	Howell et al. [92]	AAM	CD-HMM	RM-3000	Sentences	75.58%
2016	Hu et al. [94]	RTMRBM	SVM	AVLetters	Alphabet	64.63%
		RTMRBM	SVM	AVLetters2	Alphabet	31.21%
2016	Lee et al. [117]	DCT+PCA	HMM	OuluVS2	Phrases	63.00%
		RAW	PLVM	OuluVS2	Phrases	73.00%
		DCT+HiLDA	HMM	OuluVS2	Phrases	74.00%
		CNN	LSTM	OuluVS2	Phrases	81.10%
2016	Petridis et al. [173]	DBNF+DCT	LSTM	AVLetters	Alphabet	58.10%
		DBNF+DCT	LSTM	OuluVS	Phrases	81.80%
2016	Rekik et al. [189]	HOG+MBH	SVM	CUAVE	Digits	70.10%
		HOG+MBH	K-NN	MIRACL-VC	Phrases	58.10%
		HOG+MBH	SVM	OuluVS	Phrases	68.30%
		HOG+MBH	HMM	MIRACL-VC	Phrases	69.60%
		HOG+MBH	SVM	MIRACL-VC	Phrases	79.20%
2016	Saitoh et al. [196]	CFI+NIN		OuluVS2	Phrases	81.10%
		CFI+AlexNet		OuluVS2	Phrases	82.80%
		CFI+GoogLeNet		OuluVS2	Phrases	85.60%
2016	Takashima et al. [215]	CBN	HMM	ATR	Words	51.00%
2016	Wand et al. [228]	Eigenlips	SVM	GRID	Phrases	69.50% [†]
		HOG	SVM	GRID	Phrases	71.20% [†]
		Feed-Forward	LSTM	GRID	Phrases	79.50% [†]
2016	Wu et al. [238]	SDF+STLF	SVM	OuluVS2	Phrases	87.55%
2016	Zimmermann et al. [258]	PCA _{NN} +LSTM	HMM	OuluVS2	Phrases	73.00%
2017	Bear et al. [16]	AAM	HMM	AVLetters2	Alphabet	36.53% ^C [†]
		AAM	HMM	LILiR	Sentences	41.53% ^C [†]
2017	Chung et al. [44]	CNN	LSTM+Attention	OuluVS2	Phrases	91.10%
		CNN	LSTM+Attention	MV-LRS	Sentences	43.60%
2017	Chung et al. [41]	CNN	LSTM+Attention	LRW	Words	76.20%
		CNN	LSTM+Attention	GRID	Phrases	97.00%
		CNN	LSTM+Attention	LRS	Sentences	49.80%
2017	Fernandez et al. [61]	DCT+SIFT+LDA	HMM	VLRf	Sentences	20.00%
2017	Fernandez et al. [63]	DCT+SIFT+LDA	HMM	AV@CAR	Sentences	23.00%
2017	Petridis et al. [172]	Autoencoder	LSTM	OuluVS2	Phrases	84.50%
2017	Petridis et al. [176]	Autoencoder	Bi-LSTM	OuluVS2	Phrases	91.80%
2017	Petridis et al. [177]	Autoencoder	Bi-LSTM	OuluVS2	Phrases	94.70%

* V: Viseme accuracy, P: Phoneme accuracy, C: Correctness.

[†] Speaker dependent.

Table 2.6: ALR systems from 2007 to 2017 - Part III

Year	Reference	Model		Database	Recognition task	WRR (%)
		Features	Classifier			
2017	Stafylakis et al. [205]	3D-CNN+ResNet	Bi-LSTM	LRW	Words	83.00%
2017	Sterpu et al. [206]	DCT	HMM	TCD-TIMIT	Sentences	31.59% ^{V†}
2017	Sui et al. [211]	CHAVF	SVM	OuluVS	Phrases	68.90% [†]
		CHAVF	HMM	AusTalk	Digits	69.18%
2017	Thangthai et al. [219]	PCA+LDA+MLLT	DNN-HMM	TCD-TIMIT	Sentences	43.61%
2017	Thangthai et al. [218]	Eigenlips	DNN-HMM	TCD-TIMIT	Sentences	42.97%
2017	Wand et al. [229]	Feed-Forward	LSTM	GRID	Phrases	42.40%
2017	Rahmani et al. [187]	Autoencoder	DNN-HMM	CUAVE	Digits	64.9% ^P
2018	koumparoulis et al. [105]	Autoencoder	TDNN+LSTM	OuluVS2	Phrases	90.00%
2018	Fung et al. [71]	3D-CNN	Bi-LSTM	OuluVS2	Phrases	87.60%
2018	Petridis et al. [175]	3D-CNN+ResNet	Bi-GRU	LRW	Words	82.00% [†]
2018	Petridis et al. [174]	Autoencoder	Bi-LSTM	AV Digits	Phrases	69.70%
					Digits	68.00%
2018	Wand et al. [230]	Feed-Forward	LSTM	GRID	Phrases	84.70%
2018	Xu et al. [242]	3D-CNN+highway	Bi-GRU+Attention	GRID	Phrases	97.10%
2018	Afouras et al. [4]	3D-CNN+ResNet	Bi-LSTM+LM	LRS2	Sentences	37.80%
			Depthwise-CNN+LM			45.00%
			Transformer+LM			50.00%
2018	Afouras et al. [2]	3D/2DResnet+TM	TM-seq2seq + LM	LRS2	Sentences	50.0%
			TM-CTC + LM	LRS2	Sentences	45.3%
			TM-seq2seq + LM	LRS3	Sentences	42.1%
			TM-CTC + LM	LRS3	Sentences	38.2%
2019	Zhao et al. [252]	CSSMCM		CMLR	Sentences	67.52 ^Y
2019	Kandala et al. [101]	3D-CNN+2D-CNN	BiLSTM	GRID	Phrases	92.7%
				Korean	Phrases	95.8%
2019	Wand [231]	ST-CNN	BiLSTM+Attention	LRW	Words	83.34%
				LRW-1000	Words	36.91%
2019	Yang et al. [244]	LipNet		LRW	Words	83.0%
		LipNet		LRW-1000	Words	38.19%
2019	Weng et al. [234]	3D-CNN	BiLSTM	LRW	Words	84.07%
2019	Qu et al. [185]	CNN+BiGRU+FC		GRID	Phrases	97.47%
2019	Shillingford et al. [203]	ST-CNN	BiLSTM	LSVSR	Sentences	60.1%
				LRS3	Sentences	44.9%
2019	Makino et al. [134]	RNN-T		YT31k	Sentences	51.5%
		RNN-T		LRS3	Sentences	66.4%
2019	Koumparoulis et al. [106]	MobiLipNetV2	TDNN+WFST+LM	TCD-TIMIT	Sentences	43.03%
2019	Zhang et al. [249]	3D-CNN+2D-CNN+ResNet-18	STFM+CNN-seq2seq	GRID	Phrases	98.7%
				LRW	Words	83.7%
				LRS2	Sentences	48.3%
				LRS3	Sentences	36.9%
2020	Petridis et al. [178]	Autoencoder	Bi-LSTM	OuluVS2	Phrases	95.6%
				CUAVE	Digits	88.4%
				AVLetters	Letters	69.2%
				AVLetters2	Letters	42.6%
2020	Zhao et al. [253]	LIBS		CMLR	Sentences	68.73% ^Y
		LIBS		LRS2	Sentences	34.71%
2020	Luo et al. [131]	3D-CNN+ResNet-18+BiGRU	GRU	GRID	Phrases	87.7%
				LRW	Words	77.3%
				LRW-1000	Words	33.1%
2020	Chen et al. [35]	Denset+ResBi-LSTM		NSTDB	Sentences	49.56%
2020	Xiao [240]	DFTN		LRW	Words	84.13%
		DFTN		LRW-1000	Words	41.93%
2020	Zhao [251]	GLMIM		LRW	Words	84.41%
		GLMIM		LRW-1000	Words	38.79%
2020	Chen [34]	LipResNet		TCD-TIMIT	Sentences	53.8 ^P %
		LipNet		GRID	Phrases	96.9%
2020	Martinez et al. [136]	TCN		LRW	Words	85.3%
		TCN		LRW-1000	Words	41.4%
2020	Cheng et al. [36]	3D-CNN+ResNet	BiGRU	LRW	Words	83.20%
				LRS2	Words	59.60%
2020	Afouras et al. [6]	ResNet+Jasper-lip+CTC+KD		LRS2	Sentences	48.7%
		ResNet+Jasper-lip+CTC+KD		LRS3	Sentences	40.2%

* ^V: Viseme accuracy, ^P: Phoneme accuracy, ^C: Correctness, ^Y: Character.[†] Speaker dependent.

audio-based and video-based speech recognition systems. Initially, hybrid ASR systems combining traditional blocks with DNNs were proposed. More precisely, neural networks were first considered as feature extractors, mainly in combination with HMM-based classifiers. Afterward, recurrent networks, e.g. Long-Short Term Memory (LSTM) networks [77], were introduced as a suitable replacement for HMMs. More recently, end-to-end DNNs have been used to fully replace all building blocks of ASR systems by neural networks, achieving considerably higher performance than traditional systems [81, 82, 85].

A similar progression is observed for video-based systems. In Tables 2.4, 2.5 and 2.6 we see that hybrid ALR systems, firstly proposed in 2011, consist of combinations of traditional features or classifiers with neural networks [173, 210, 154, 157, 158]. In subsequent years, there has been a tendency toward ALR systems based purely on DL, known as end-to-end DNN architectures.

In this section, the DNN-based systems presented in Tables 2.4, 2.5 and 2.6 are analyzed. Similarly to Section 2.2.1, we firstly explain the architectures' components and then compare the different systems in terms of performance.

Configuration of DNN-architectures

ALR systems based on end-to-end DNNs follow a similar pipeline to traditional ones (shown in Section 2.2.1-Figure 2.3). Similarly to the previous section, we will compare systems in terms of feature extraction and classification stages.

We start by showing in Figures 2.8 and 2.9 how frequently the different types of DNNs have been integrated into ALR systems as a feature or classification technique. In Figure 2.8 we observe that Convolution Neural Networks (CNN) have been the most used networks to extract features, but other DNNs such as Feed-forward networks or Deep Belief Networks (DBN) have also been used. In terms of classifiers, in Figure 2.9 we can see a predominance of LSTMs, although CNNs, Feed-forward DNNs and DBNs have also been used.

Looking at Tables 2.4, 2.5 and 2.6 we observe that there are 24 end-to-end DL architectures, from which 11 consist of combinations of CNNs and RNNs (LSTMs or GRUs). Thus, this combination stands out as the most used DL architecture for ALR and we will analyze it in more detail. In Figure 2.10 we show a CNN-LSTM baseline system where a sequence of video frames are processed by a convolutional network followed by a recurrent network. CNNs have been established as a powerful model to extract visual features for image recognition and classification tasks [107, 214] and consist of alternating convolutional layers and pooling layers. The convolutional layers compute the inner product between linear filter and the receptive field and then they are followed by a non-linear activation function (e.g. sigmoid, tanh, ReLU). On the other hand, LSTMs are recurrent neural networks (RNN) useful for modeling

sequences due to their cyclic connections that form a temporal memory [161, 83]. LSTMs have been widely used because they solve the vanishing and exploding gradient problem [20] that appears in conventional RNNs. In contrast to RNNs, LSTMs have a cell unit that is regulated by 3 gates, known as input, output and forget gates, which use additive and multiplicative connections to ensure constant error flow, thus retaining short- and long-context information.

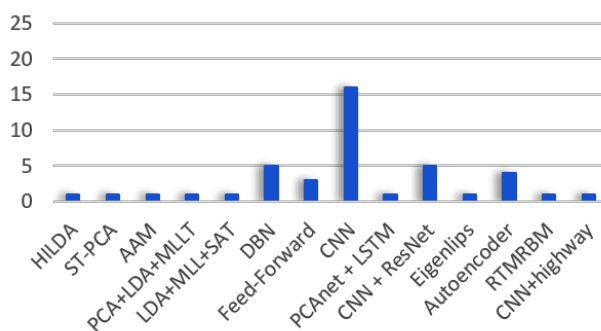


Figure 2.8: DNN-based systems. Number of times that each feature technique has been used from 2007 to 2018.

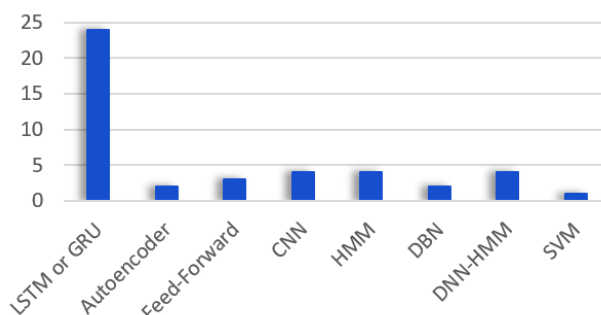


Figure 2.9: DNN-based systems. Number of times that each classification method has been used from 2007 to 2018.

Architectures based on CNNs and LSTMs

Several authors have proposed CNN-LSTM networks that follow the baseline in Figure 2.10. For instance, Chung et al. [43] proposed a network that performs sentence-level classification. Notice that "sentence-level classification" means that the system's output is restricted to a finite number of possible sentences, which therefore act as the classes of a classification problem. The architecture

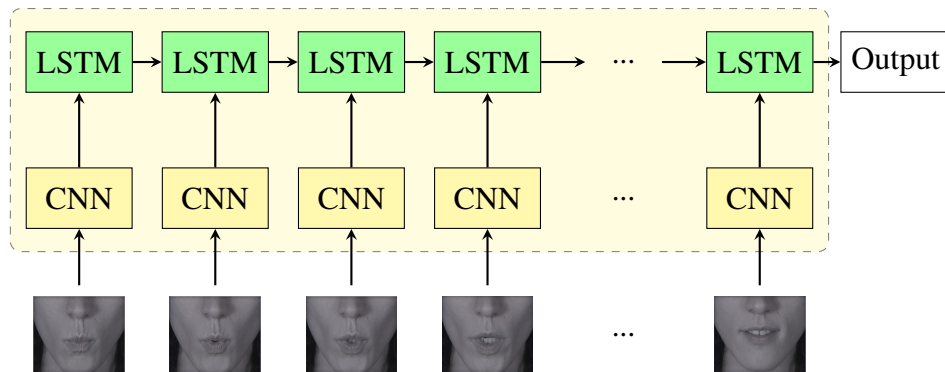


Figure 2.10: Baseline DL architecture for lip-reading, consisting of combinations of CNNs and LSTMs.

inputs gray-scale images into a convolutional network, named SyncNet, which consists of five convolutional layers followed by two fully-connected layers. For each frame, the output of the last CNN layer is the input to a single-LSTM layer that accumulates the contribution of each frame and returns the estimated class at the end of the sequence. The block diagram of this architecture is provided in Fig. 2.11-(a). Still within the same work [43], Chung et al. compare the proposed CNN with a pre-trained network, known as VGG-M (Fig. 2.11-(b)). VGG-M consists of five convolutional layers followed by three fully-connected layers pre-trained in the ImageNet database [107]. The VGG-M output is the input to a single LSTM layer that performs the classification at the end of the sequence, similarly to SyncNet. As we will see in Section 2.2.2, in spite of having an additional fully connected layer, the pre-trained VGG-M did not perform as good as SyncNet given that the training of the latter was much more specific to the lip-reading task.

Lee et al. [117] proposed a DNN architecture that performs sentence-level classification (Fig. 2.12-(a)). Their system inputs RGB normalized images that are processed by a CNN with two convolutional layers and one fully-connected layer. They also define a temporal model based on two LSTM layers that receive the CNN features and accumulate the contribution of each frame until the end of the sequence, which is finally processed by a fully connected layer that returns the classification of the whole sequence into a phrase.

Assael et al. [11] proposed LIPNET, an end-to-end DL-architecture that also performs sentence-level classification (Fig. 2.12-(b)). The model's input is a fixed-length sequence of RGB normalized images that are processed by three spatio-temporal convolutional layers. The output features of the CNN are fed to two Bidirectional Gated Recurrent Network (GRU) layers that are finally

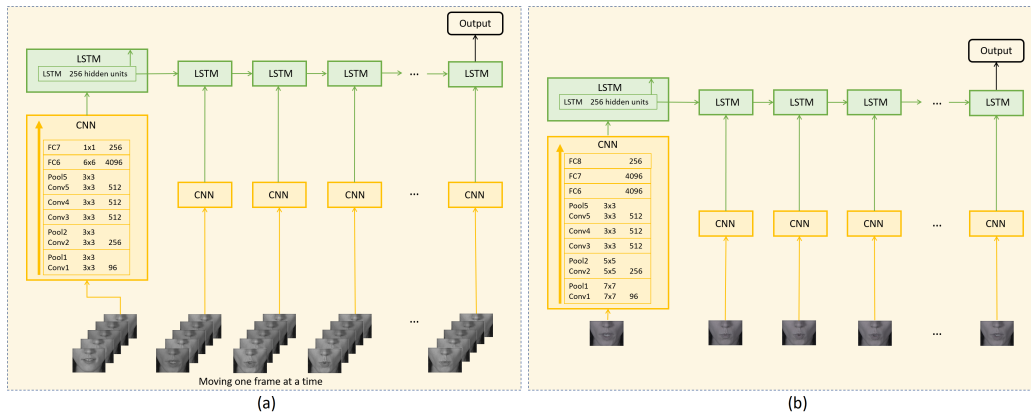


Figure 2.11: Architectures from [43]. (a) Combination of SyncNet and LSTMs; (b) Combination of VGG-M and LSTMs

followed by a linear transformation at each time-step and a softmax over the vocabulary (which in this case is a character-based representation). This end-to-end model is trained with a Connectionist Temporal Classification (CTC) [81] network that has a softmax output layer with as many units as the number of labels in the vocabulary plus one unit for the blank character ”_”. The CTC computes the probability of all possible combinations of a string. For example, if the sequence length is fixed to 3, the CTC defines the probability of a string ”am” as $p(aam) + p(amm) + p(_am) + p(a_m) + p(am_)$. The model predicts frame labels and finds the optimal alignment between the predictions and the output sequence (which is a full-sentence within the possible pre-defined classes).

On the other hand, Stafylakis et al. [205] proposed a system that performs

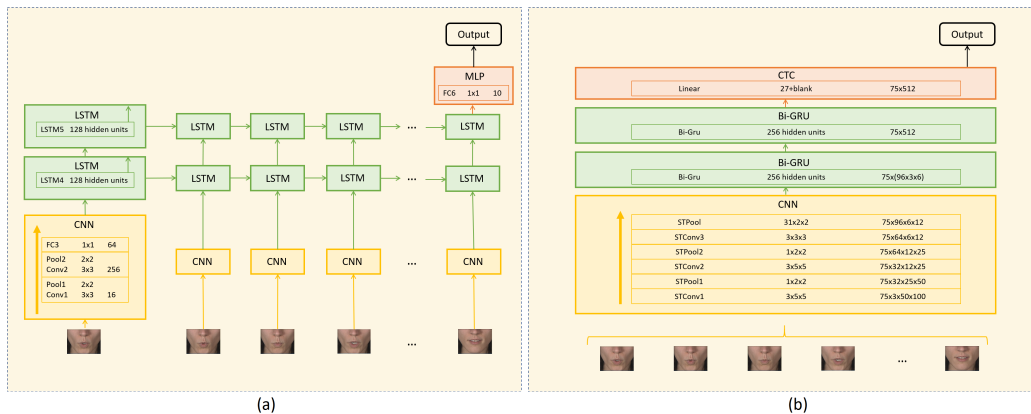


Figure 2.12: (a) Architecture from [117]; (b) Architecture from [11]

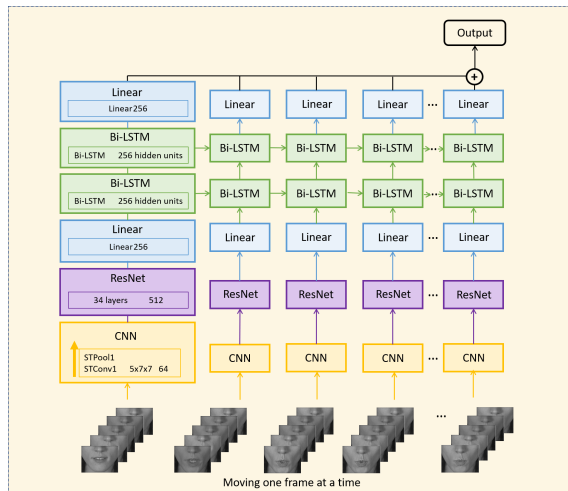


Figure 2.13: Architecture from [205]

word-level classification (Fig. 2.13). In their model, the inputs are video sequences of gray-scale normalized images, with a fixed duration of 1 second. The proposed architecture is based on a spatio-temporal convolutional layer followed by a residual network (ResNet [88]). The residual network consists of 34-layers (including convolutional, pooling and fully-connected layers) that progressively reduce the spatial dimensionality with max pooling layers, until the output becomes a single dimensional vector per time step. Then, these vectors are used as input features to two bidirectional LSTMs (Bi-LSTM) [83] (two in each direction) which are concatenated at each time step for classification. Differently from previous works, the classification is not performed at the last time step of the LSTM output, once all the sequence has been encoded by the LSTM, but the softmax is applied at each time step. Hence, the overall loss is defined as the aggregated loss over all time steps.

Notice that these two last systems [11, 205] used Bi-LSTMs or Bi-GRUs for their ability to produce outputs conditioned on past and future contexts, as opposed to the standard LSTMs that work only in one direction. Other very recent works have also explored the use of these bi-directional networks. On one hand, Petridis et al. [175] proposed a model very similar to [205], where the main difference between both lip-reading architectures is that [175] used Bi-GRU networks with a bigger number of hidden units instead of the Bi-LSTMs networks used in [205]. On the other hand, Fung et al. [71] used Bi-LSTMs for sentence-level classification. Their network consists of 8 spatiotemporal convolutional layers followed by a maxout activation function without pooling layer that is fed to the Bi-LSTM layer. The final output is obtained with a softmax layer at the last time step of the sequence.

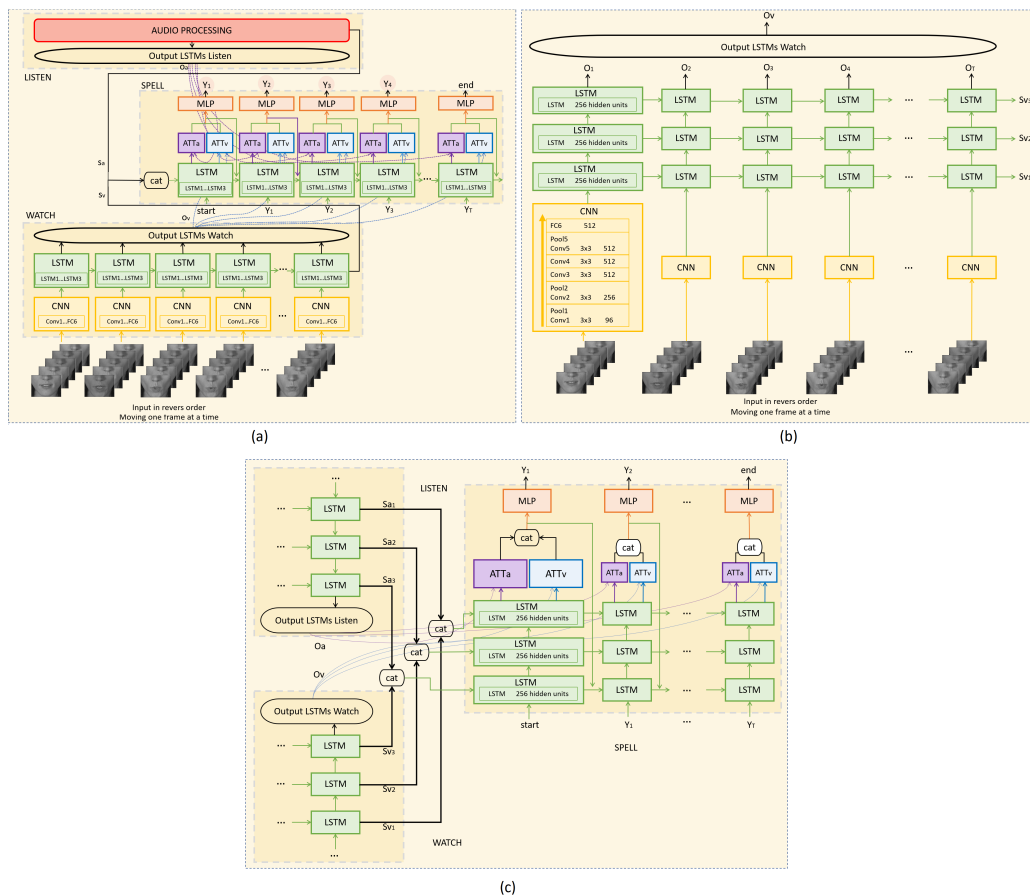


Figure 2.14: Architecture from [44] (a) WLAS; (b) WATCH; (c) SPELL

Chung et al. proposed a system for AVSR [41] and another one for ALR [44] (Fig. 2.14-(a), 2.14-(b) and 2.14-(c)). For the AVSR system, they proposed an end-to-end network based on four main modules, named *Watch*, *Listen*, *Attend and Spell*, that learned to predict characters from spoken sentences. The *Watch* module receives video input and consists of five 3D-convolutional layers followed by one fully-connected layer and then three LSTM layers stacked one behind the other to catch different levels of abstraction. A similar network is employed for *Listen* to process audio. The *Spell* module consists of three LSTMs, two attention mechanisms (for the audio and visual contexts provided by *Watch* and *Listen*) and a multi-layer perceptron (MLP). Thus, *Spell* LSTMs use: the previous character, the previous LSTM state and the concatenation of the last time-step of *Watch* and *Listen* LSTMs. Next, two context vectors are computed in the *Attend* module, from audio and visual contexts. These context vectors are computed at each time-step by the attention mechanisms. The

attention mechanisms use the output produced by the *Watch* or *Listen* LSTMs at each time step and the current outputs of *Spell* LSTMs. Finally, the probability distribution of the output character is generated by the MLP with a softmax layer over the output. The authors emphasize that gray-scale image sequences are processed in reverse time-order, as this was found to improve results. They also explain that attention is crucial for the system because without it the model forgets the input signal, and produces an output sequence that does not correlate with the input beyond the first word or two (which the model gets correct, as these are the last words to be seen by the encoder). In addition, unidirectional encoders for the *Watch* and *Listen* modules were compared with bidirectional encoders, but the latter networks took significantly longer to train, while providing no obvious performance improvement. For the ALR system proposed in [44], where audio information is not available, the same architecture was proposed except that there were no audio attention nor *Listen* blocks.

As the last example of the CNN-LSTM architecture, Xu et al. [242] presented a network named LCANet that performs character-level classification. The video encoder of LCANet has three components: 3D convolutions, a highway network, and Bi-GRU networks. LCANet feeds 3 consecutive frames into a 3D convolutional neural network to encode both visual and short temporal information. Then, they stack two layers of highway networks [204] on top of the 3D-CNN. The highway network module has a pair of transform gate and carries a gate that allows the deep neural network to carry some input information directly to the output. These networks have been enabled to encode much richer semantic features. At the end of the video encoding, Bi-GRU networks are feed after the highway networks to encode long-term temporal information. To capture information explicitly from a longer context, LCANet feeds the encoded spatiotemporal features into a cascaded attention-CTC decoder. The attention mechanism debilitates the constraint of the conditional independence assumption in CTC loss, but it improves the modeling capability on the lipreading problem and can give better predictions on visually similar visemes.

Other DL-architectures

Some authors have also proposed end-to-end architectures that do not follow the CNN-LSTM baseline from Figure 2.10. For instance, Wand et al. proposed three DNN architectures [228, 229, 230] that perform word-level classification. The system proposed in [228] (Fig. 2.15-(a)) consists of one feed-forward layer followed by two LSTMs and a softmax layer to perform classification within a set of pre-defined classes. Similarly, the system proposed in [229] (Fig. 2.15-(b)) consists of three feed-forward layers followed by one LSTM layer and a softmax layer to perform classification within the set of words. In order to mitigate the

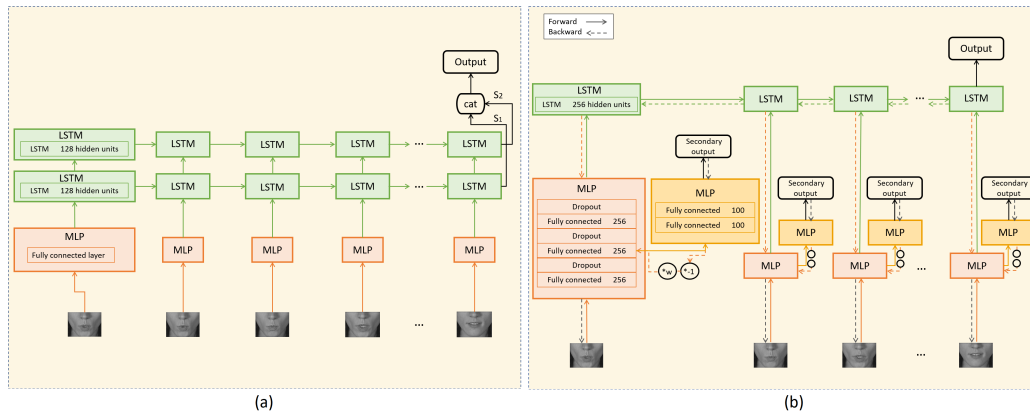


Figure 2.15: (a) Architecture from [228]; (b) Architecture from [229]

discrepancy between known and unknown speakers, it incorporates domain adversarial training, by means of an intermediate layer driven to learn a domain-agnostic representation of the input data. Specifically, at the second feed-forward layer, a supplementary network consisting of two feed-forward layers and a softmax layer is integrated to perform speaker classification. The incorporation of the adversarial network is supposed to be beneficial because by feeding its inverted gradient into the main network, the system is prevented from learning speaker-dependent features. Finally, the system proposed in [230] consists of three feed-forward layers followed by one LSTM layer and a softmax layer that performs word classification at the end of the sequence. In this architecture all layers, including the LSTM, have the same number of neurons.

Chung et al. [42] also proposed a DNN architecture that performs word-level classification (Fig. 2.16-(a)). The method pre-processes each input frame with a first convolutional layer whose outputs are concatenated so that the whole sequence is sent to a second convolutional layer. The output of the second layer is fed into the following layers, which have a similar structure to VGG-M: three additional convolutional layers, three fully connected layers and one softmax layer.

Saitoh et al. [196] proposed an end-to-end system for sentence-level classification that instead of processing the sequence frame by frame, constructs a macro image by concatenating a subset of the whole video sequence, which they call concatenated frame image (CFI). They test the CFI in combination with three pre-trained CNNs: Networks in Networks (NIN) [120], AlexNet [107] and GoogLeNet [214]. NIN is a novel network that replaces the usual linear convolutional layers with MLP-Convolutional layers (mlpconv). Specifically, Saitoh et al. used four mlpconv followed by a spatial max pooling layer. AlexNet consists of five convolutional layers followed by three fully connected layers, and

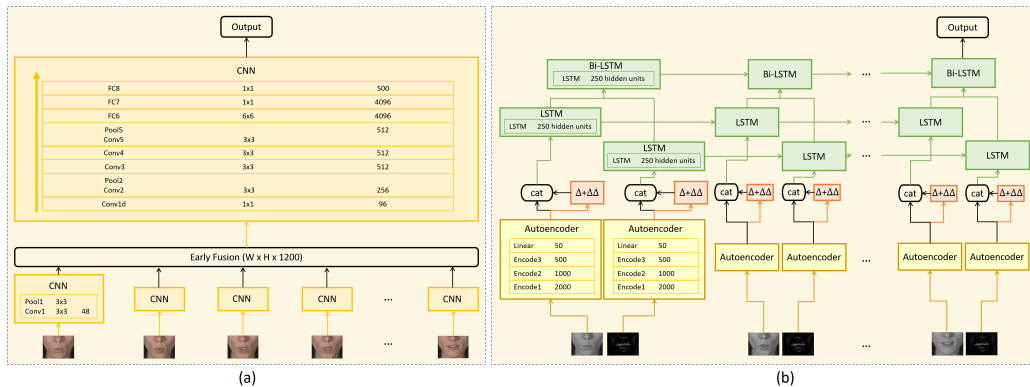


Figure 2.16: (a) Architecture from [42]; (b) Architecture from [172]

GoogleLeNet is a twenty-two layer deep network that uses a sparsely connected architecture (inception modules) to avoid computational bottlenecks. Despite the different architectures of the three networks, their performance in the ALR tests reported by Saitoh et al. [196] were fairly similar, with differences that did not exceed 5% WRR between them.

Petridis et al. [172, 176, 177, 174] proposed four end-to-end systems for sentence-level classification. Firstly in [172] (Fig. 2.16-(b)), they proposed a system based on two independent streams; the first one extracts features directly from single-images, while the second one extracts features from the difference between two consecutive frames. Both streams follow a bottleneck architecture with three hidden layers and one linear layer. At the end of the bottleneck architecture, the first and second derivatives are computed and appended to the bottleneck layer. The output of the bottleneck network of each stream is fed into an LSTM layer. Finally, the LSTM outputs of both streams are concatenated and fed into a Bi-LSTM in order to fuse their information. The output layer is a softmax layer that performs the classification using the last time step of the Bi-LSTM output, once all the sequence has been encoded. On the other hand, the system proposed by Petridis et al. in [176] is a very similar network that also incorporates audio input. Specifically, the frame difference data is replaced by audio features, so that one stream per modality is used. They also replace the LSTM networks at the end of each stream by Bi-LSTMs. The third system presented by Petridis et al. in [177] tackled multi-view lip-reading for sentence-level classification. It consists of three identical streams which extract features from three images captured from different view angles. The streams follow the same architecture from [176] and their outputs are concatenated and fed into a Bi-LSTM and a softmax layer that performs the classification similarly to the two architectures previously described. Finally, the fourth system [174] was proposed as a modification of [176]. The key difference is that the new

system used only a single stream (corresponding to the video frames) instead of the use of two streams proposed previously.

A transfer DL framework was presented by Moon et al. [147] for alphabet recognition. The system uses audio and visual information independently to learn abstract representations of the data using a standard deep belief network (DBN) with multiple Restricted Boltzmann machines (RBMs). This allows for semantic-level transfer between the source and target modules. Both DBNs, for audio and visual information are built with the same number of intermediate layers, and then inter-modal embeddings are learned for each layer. Then, the learned mappings between the source and target are used to fine-tune the network with the transferred data and categorize each sequence into a letter.

More recently, Afouras et al. [4] proposed three systems that perform character-level classification. The visual front-end is common across the three systems and consists of a 3D CNN on the input image sequence, with a filter width of five frames, followed by a ResNet which gradually decreases the spatial dimensions as depth increases. In contrast, the temporal back-end that receives the frame feature vectors and outputs a sentence character by character, is different for each system. The first one consists of three stacked Bi-LSTMs trained with CTC loss and decoding is performed with a beam search that incorporates prior information from an external language model. The second system uses depth-wise separable convolution layers, which consist of a separate convolution along the time dimension for every channel followed by a projection along the channel dimensions. The network contains 15 convolutional layers that were trained with a CTC loss and decoding is performed as described in the same way as the previous system. Finally, the last system has an encoder-decoder structure based on multi-head attention layers. It uses a base model with 6 encoder and decoder layers and 8 attention heads. This system has been trained with cross-entropy loss instead of CTC, hence it would be expected to implicitly learn an internal language model. Nevertheless, authors report that integrating an external language model in the decoding process improved their results.

Performance comparison

In this section we compare the performance of both hybrid and end-to-end DNN-based architectures. We compare the methods from Tables 2.4, 2.5 and 2.6 that have been evaluated in the most common databases, being them AVLetters, GRID, LRW and OuluVS2.

For alphabet recognition, we find four DNN-based systems evaluated in the well known AVLetters database [154, 147, 173, 94]. The first one was presented by Ngiam et al. [154] and consists of PCA features followed by a deep autoencoder, obtaining a classification accuracy of 64.40% WRR. In contrast,

Moon et al. [147] proposed a method to obtain abstract representations of the raw data using a standard DBN. They fine-tune the video model with additional information transferred from audio data, obtaining 55.30% WRR. Petridis et al. [173] proposed to first train a deep autoencoder to compress the high dimensional image data into a low dimensional representation (named bottleneck features). Next, DCT features are computed to complement bottleneck ones and fed to an LSTM network to model the temporal dynamics, obtaining 58.10% WRR. Finally, Hu et al. [94] proposed a system based on multimodal RBMs (MRBMs), named Recurrent Temporal Multimodal Restricted Boltzmann Machines (RTMRBMs), which have the ability to extract semantic information from multisensory data and learn a joint representation across audiovisual modalities. They reported 64.63% WRR. Interestingly, these results are below those obtained by some traditional systems, e.g the RFMA-based system presented in [171] obtained 69.60% WRR. Thus, for letter recognition in datasets such as AVLetters, traditional systems still outperform DL-systems. The reason for this seems related to the dataset size, which is not large enough to train robust DL systems.

For word or sentence recognition, the most used databases have been GRID, LRW and OuluVS2. For the GRID corpus, we found six different architectures. Wand et al. presented three models for this database: the first one [228] consists of one Feed-forward layer followed by two recurrent LSTM layers and reported 79.50% WRR, while the second and third systems [229], [230] combine three Feed-forward layers with an LSTM layer and reported 83.30% and 84.70% WRR for speaker-dependent experiments and 42.40% WRR in [229] for experiments in which the test speakers were unknown to the system. In contrast, Assael et al. [11] proposed a spatio-temporal CNN in combination with Bi-LSTMs and obtained a higher recognition rate of 93.40% WRR. Chung et al. [41] obtained 97.00% WRR with a system based on CNN and LSTM networks combined with attention mechanisms. Finally, Xu et al. [242] outperformed previous methods with a system that combines 3D-CNNs, highway networks, Bi-GRUs and attention mechanisms, obtaining slightly higher performance than [41] with 97.10% WRR. There is a considerable improvement in performance with respect to traditional systems, where the highest accuracy was 57.00% WRR reported by [104].

For the LRW database, Chung et al. [42] presented an end-to-end architecture based on CNNs, reporting 61.10% WRR. Stafylakis et al. [205] presented a system based on 3D-CNN, residual networks and Bi-LSTMs and reported more than 20% improvement (83.00% WRR). Similarly, Petridis et al. [175] presented a system based on 3D-CNN, residual networks and Bi-GRU networks and reported 82.00% WRR. In yet another contribution, Chung et al. [41] proposed a system based on CNN and LSTM networks combined with attention mechanisms

and obtained the best results reported so far, with 84.50% WRR.

For the OuluVS2 dataset, 13 architectures have been presented. Saitoh et al. [196] and Chung et al. [42] presented several end-to-end systems mainly based on CNNs. The three systems proposed by Saitoh et al. reported recognition rates between 81.10% and 86.50% WRR, while Chung et al. reported 94.10% WRR. The main difference between these two works is that the networks in [196] used CFIs as input while [42] used directly a single image. In addition, Saitoh et al. used three well known pre-trained models based on CNNs: NIN [120], AlexNet [107] and GoogLeNet [214], while Chung et al. trained the network from scratch for the specific task of lip-reading. Several architectures were also proposed with LSTMs or Bi-LSTMs as classifiers. For these systems, different models to extract features were applied: CNNs in [117, 44], VGG-M and SyncNet in [43], autoencoders in [172, 176, 177], 3D-CNN in [71] and PCA-NN in [258]. The latter one, in addition, used HMMs to model the temporal dynamics. For these architectures, the reported recognition rates were between 31.90% and 94.70% WRR. The lowest recognition rate corresponds to the system using VGG-M [43]. This comparatively low accuracy can be explained because VGG-M was pre-trained on ImageNet, a large database for object recognition and classification tasks, but not specific for lip-reading. In contrast, Petridis et al. [177] presented a system based on encoded features that reported the highest performance of 94.70% WRR, nearly followed by Chung et al. [42] with 94.10% WRR. Nevertheless, compared to traditional architectures, there is a significant improvement of at least a 20% with respect to the highest performing traditional system, achieving 74.00% WRR in [117].

From the above paragraphs we can see that DNNs brought substantial accuracy improvements to ALR systems on databases such as GRID or OuluVS2, which focus on word- or sentence- classification tasks. These improvements have encouraged researchers to address more realistic settings and propose systems that target continuous lip-reading. Such settings are considerably more challenging than those found in word- or sentence-classification tasks, because each sentence has an unknown structure and can contain an arbitrary number of words whose time-boundaries are not known beforehand. For these reasons, when targeting continuous lip-reading it is convenient to predict smaller structures that approach the minimum distinguishable language units. Recent advances in end-to-end DL architectures have indeed focused on ALR systems that try to predict phonemes [215, 158, 157, 148] or characters [41, 44, 4, 242], instead of full words or pre-defined sentences. For example, Mroueh et al. [148] proposed Feed-forward DNNs to predict phonemes using the IBM AVSR database, a large scale non-public AV database. Other architectures using CNNs and HMMs were presented by Noda et al. [158, 157] and by Takashima et al. [215]. They tried to

recognize Japanese phonemes using the ATR Japanese corpus [109] and obtained 22.50% WRR, 37.00% WRR and 51.00% WRR, respectively. Another architecture evaluated in the highly used GRID corpus has been recently presented by Xu et al. [242] for character-based classification. This very deep network combines 3D-CNNs, highway networks, Bi-GRUs and attention mechanisms and reported 97.10% WRR. In contrast, Chung et al. [41, 44] presented an architecture based on CNN and LSTM networks combined with attention mechanisms. They evaluated their system in recently recorded large-scale databases such as MV-LRS and LRS, obtaining for character-based recognition 43.60% WRR and 49.80% WRR, respectively for each dataset. More recently, Afouras et al. [4] presented a comparison of three architectures dealing with character-based recognition evaluated on the LRS dataset. The architectures share the same visual features and only differ in the sequence classification; they obtained 37.80% WRR for the model using Bi-LSTMs, 45.00% WRR for the one using depth-wise convolutional layers and 50.00% WRR for the one using encoder-decoder with multi-head attention layers.

Thus, most recent DNN-based architectures report WRRs that, despite the different experimental settings, nearly double the performance reported by traditional systems, with WRRs of about 20% [63, 61, 113]. While this constitutes a great step forward in continuous lip-reading, it is worth noting that these results are still far from a system that can fully decode visual speech. Indeed, in real-world scenarios, the top-performing ALR systems currently approach WRRs of 50%, which means that we cannot recognize about half of the message. Thus, DNN-based systems and large-scale databases have significantly advanced the field but continuous ALR remains still an open problem.

2.3 Summary and Conclusions

In this survey, we review the progression of ALR systems from 2007 to 2017 which highlights the technology shift from traditional architectures, typically consisting of image features in combination with HMMs, toward end-to-end DNN architectures, currently dominated by CNN-features in combination with LSTMs.

In both the traditional and the DNN-based systems, we can conceptually identify two major blocks specific to ALR whose objectives are: i) to parametrize the visual information observable at a given time instant or window, and ii) to map the visual features into speech units while incorporating temporal context, i.e. constraints to ensure that the decoded message is coherent. The latter provides robustness against noisy or imperfect estimates from the visual cues and helps to disambiguate between visually similar speech units.

Traditional ALR systems mainly consist of features based on appearance or image transforms in combination with HMMs that model the temporal dynamics of the spoken sequence using short term context information. While HMMs can be considered the de-facto standard for modeling context, a variety of features have been explored with the goal to find the best descriptor for visual speech. As shown in Section 2.2.1, the most widely used features in visual-speech systems have been DCT and AAMs, but there is no agreement on which feature would be optimal.

In the last years, we observe how DNN-based systems have quickly started to replace all the blocks from traditional systems by end-to-end DNNs. In this survey, we discuss the most popular DNN architectures for ALR systems and compare several variations that follow the same baseline structure (i.e. combinations of CNNs and LSTMs). In particular, variants on the feature side include different types of data used to feed the CNNs (e.g. RGB or gray-scale images, 3D or 2D structures), and network specifications (e.g. number of convolutional and fully-connected layers). In terms of classification, ALR researchers have explored LSTM networks that differ in how the output is decoded (e.g. step by step or at the end of the sequence), the network's direction (forward, backward or bidirectional), and the number of layers (which relates to the context scale that is considered). In addition, we comment other DNNs used for lip-reading that explore alternatives to the CNN-LSTM baseline, such as Feed-Forward networks, DBN, or CNNs.

Comparing traditional systems with DL architectures we observe that the latter provide a significant improvement in terms of performance. For instance, for the GRID corpus, several DL architectures considerably outperformed the best traditional system with up to a 40% improvement, e.g. Assael et al. [11], Chung et al. [41] and Xu et al. [242] proposed end-to-end architectures that achieved up to 97% WRR, compared to the 57% WRR obtained by Kolossa et al. [104]. Similarly, in the OuluVS2 database, DNN-systems [196, 43] reported more than 20% improvement with respect to the best-performing traditional system, which achieved 74% WRR [117].

Nevertheless, the remarkable results of end-to-end DL architectures addressing word or sentence recognition in databases such as GRID or OuluVS2, cannot be directly extrapolated to more realistic settings that target continuous lip-reading. In word or sentence recognition tasks, the output of the system is restricted to a pre-defined number of possible classes, in contrast to continuous lip-reading where the target is natural speech. In this way, continuous lip-reading systems must be able to decode any word of the dictionary and process sentences that contain an arbitrary number of words with unknown time-boundaries. Thus, recent attempts to produce continuous lip-reading systems have focused on elementary language structures such as characters or phonemes. For instance,

hybrid architectures for continuous speech recognition in Japanese [158, 157, 215] have targeted phonemes achieving between 22% and 51% WRR, while Chung et al. [41] and Afouras et al. [4] achieved near 50% WRR targeting characters for a large-scale dataset in English.

Despite the recognition rates for continuous lip-reading may appear modest in comparison to those achieved by audio-based systems, the field has undeniably made a significant step forward. Interestingly, an analogous effect can be observed when humans try to decode speech: given sufficiently clean signals, most people can effortlessly decode the audio channel, but would struggle to perform lip-reading, since the ambiguity of the visual cues makes it necessary the use of further context to decode the message. Thus, it is not surprising that the main challenges in ALR systems regard to the robustness to visual ambiguities through the modeling of context information.

Most recent works suggest that the optimal modeling of temporal sequences is still an open problem, which is currently been tackled by means of recurrent neural networks. Specifically, LSTMs have been widely used for modelling sequences because of their ability to retain both short- and long-term context information in their cell structures, although it is not clear how to take full advantage of such ability. For instance, several authors have tried to model different scales of context by adding multiple LSTM layers, aiming to introduce constraints related to bigger speech structures such as connected phonemes, syllables, words or sentences. Other authors have used bidirectional networks, (widely used in audio speech recognition because of their ability to model past and future context), which should be helpful for dealing with visual ambiguities that are related to previous and posterior mouth positions (i.e. a similar idea to that from triphoneme models). However, bidirectional networks involve a higher computational cost than unidirectional ones and require that the whole signal is available beforehand, not allowing for real-time decoding. Finally, attention models have also been recently explored because they help to highlight the most relevant pieces of information from a large amount of data potentially available. Thus, current efforts tend toward techniques that allow more comprehensive modelling and interpretability of the retained context.

Chapter 3

OPTIMIZING PHONEME-TO-VISEME MAPPING FOR CONTINUOUS LIP-READING IN SPANISH

Despite the common intuition that speech is something that we hear, there is overwhelming evidence that the brain treats speech as something that we hear, see, and even feel [192]. Visual cues are often used unconsciously and to a different extent for different individuals, depending on aspects such as the hearing ability [37], or the acoustic conditions, e.g. the visual channel becomes more important in noisy environments or when someone is speaking with a heavy foreign accent [54], [213], [89], [191]. Furthermore, the visual channel is the only source of information to understand the spoken language for people with hearing disabilities [200], [183], [10].

In the literature, much of the research has focused on Automatic Speech Recognition (ASR) systems, treating speech primarily as an acoustic form of communication. Currently, ASR systems are able to recognize speech with very

Adapted from: Fernandez-Lopez, A., & Sukno, F. M. (2019). Optimizing Phoneme-to-Viseme Mapping for Continuous Lip-Reading in Spanish. In Cláudio, A. P., Bechmann, D., Richard, P., Yamaguchi, T., Linsen, L., Telea, A., Imai, F., and Tremeau, A. (Ed.). *Computer Vision, Imaging and Computer Graphics - Theory and Applications*, (pp. 305-328). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-12209-6_15.

Fernandez-Lopez, A., & Sukno, F. M. (2017). Automatic viseme vocabulary construction to enhance continuous lip-reading. In Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)-Volume 5: VISAPP; 2017 Feb 27-Mar 1; Porto, Portugal. Setúbal, Portugal: SCITEPRESS, 2017. p. 52-63.. SCITEPRESS. DOI: 10.5220/0006102100520063.

high accuracy when the acoustic signal is not degraded. However, when the acoustic signal is corrupted, the performance of ASR drops and there is the need to rely also on the information provided by the visual channel, which relates to the movement of the lips, teeth, tongue, and other facial features. This has led to research in Audio-Visual Speech Recognition (AVSR) systems, which try to balance the contribution of the audio and the visual information channels to develop systems that are robust to audio artifacts and noise. AVSR systems have been shown to significantly improve the recognition performance of audio-based systems under adverse acoustic conditions [183], [51].

On the other hand, in the last decades there has been an increased interest in decoding speech exclusively using visual cues, leading to Automatic Lip-Reading (ALR) systems [51], [151], [255], [246], [210], [41], [173], [7], [42], [228]. Nonetheless, ALR systems are still behind in performance compared to audio- or audio-visual systems. This can be partially explained by the greater challenges associated with decoding speech through the visual channel, when compared to the audio channel. Specifically, one of the key limitations in ALR systems resides on the visual ambiguities that arise at the word level due to homophemes, i.e characters that are easily confused because they produce the same or very similar lip movements.

Keeping in mind that the main objective of speech recognition systems is to understand language, which is structured in terms of sentences, words and characters, going from larger to smaller speech entities. More precisely, the standard minimum unit in speech processing is the *phoneme*, defined as the minimum distinguishable sound that is able to change the meaning of a word [222]. Similarly, when analyzing visual information many researchers use the *viseme*, which is defined as the minimum distinguishable speech unit in the video domain [66]. However, due to visual ambiguities the correspondence between both units is no one-to-one and not all the phonemes that are heard can be distinguished by observing the lips. There are two main types of ambiguities: *i*) there are phonemes that are easily confused because they are perceived visually similar to others. For example, the phones /p/ and /b/ are visually indistinguishable because voicing occurs at the glottis, which is not visible. *ii*) there are phonemes whose visual appearance can change (or even disappear) depending on the context (co-articulated consonants). This is the case of the *velars*, consonants articulated with the back part of the tongue against the soft palate (e.g: /k/ or /g/), because they change their position in the palate depending on the previous or following phoneme [146]. Thus, visemes have usually been defined as the grouping of phonemes sharing the same visual appearance [130], [194], [32]. Nonetheless, there is no consensus on the precise definition of the different visemes nor on their number, or even on their usefulness [32], [66], [44], [194]. There are discrepancies on whether there is more information in the

position of the lips or in their movement [130], [194], [32] and if visemes are better defined in terms of articulatory gestures (such as lips closing together, jaw movement, teeth exposure) which relates the use of visemes as a form of model clustering that allows visually similar phonetic events to share a common model [32], [66], [89].

Then, when designing ALR systems, one of the most important challenges is how to make the system robust to visual ambiguities. Consequently several different viseme vocabularies have been proposed in the literature typically with lengths between 11 and 15 visemes [18], [87], [183], [152]. For instance, Goldschen et al. [79] trained an initial set of 56 phones and clustered them into 35 visemes using the Average Linkage hierarchical clustering algorithm. Jeffers and Barley [99] defined a phoneme-to-viseme mapping from 50 phonemes to 11 visemes in the English language (11 visemes plus *Silence*). Neti et al. [152] investigated the design of context questions based on decision trees to reveal similar linguistic context behavior between phonemes that belong to the same viseme. For the study, based on linguistic properties, they determined seven consonant visemes (bilabial, labiodental, dental, palato-velar, palatal, velar, and two alveolars), four vowels, an alveolar-semivowel and one silence viseme (13 visemes in total). Bozkurt et al. [27] proposed a phoneme-to-viseme mapping from 46 American English phones to 16 visemes to achieve natural looking lip animation. They mapped phonetic sequences to viseme sequences before animating the lips of 3D head models. Ezzat and Poggio [60] presented a text-to-audiovisual speech synthesizer which converts input text into an audiovisual speech stream. They started grouping those phonemes which looked similar by visually comparing the viseme images. To obtain a photo-realistic talking face they proposed a phoneme-to-viseme mapping with 6 visemes that represent 24 consonant phonemes, 7 visemes that represent the 12 vowel phonemes, 2 diphthong visemes and one viseme corresponding to the silence.

Contribution: In this work, we propose to automatically construct a phoneme-to-viseme mapping based on visual similarities between phonemes to maximize word recognition. We investigate the usefulness of different phoneme-to-viseme mappings, obtaining the best results for intermediate alphabet lengths. We evaluate an ALR system based on DCT and SIFT descriptors and Hidden Markov Models (HMMs) in two Spanish corpora with continuous speech (AV@CAR and VLRf) containing 19 and 24 speakers, respectively. Our results indicate that we are able to recognize 47% (resp. 51%) of the phonemes and 23% (resp. 21%) of the words, for AV@CAR and VLRf. We also show additional results that support the usefulness of visemes. Firstly, we show qualitative results by comparing the average lip-images per subject and phoneme of several subjects from both databases, which clearly illustrate the difficulty to perceive differences between phonemes that are known to produce

visual ambiguities. Secondly, we also analyze the results by looking at the confusion matrices obtained with our system trained with and without using visemes as an intermediate representation. Experiments on a comparable ALR system trained exclusively using phonemes at all its stages confirm the existence of strong visual ambiguities between groups of phonemes. This fact and the higher word accuracy obtained when using phoneme-to-viseme mappings, justify the usefulness of visemes instead of the direct use of phonemes for ALR. This paper is an extended and revised version of a preliminary conference report that was presented in [63].

3.1 ALR system

ALR systems typically aim at interpreting the video signal in terms of visual units, and usually consist of 3 major steps: 1) Lips localization, 2) Extraction of visual features, 3) Classification into sequences. In this section we start with a brief review of the related work and then provide a detailed explanation of our method.

3.1.1 Related Work

Much of the research on ALR has focused on digit recognition, isolated words and sentences, and only more recently in continuous speech.

For continuous speech recognition: [183] applied fast DCT to the mouth region and trained an ensemble of 100 coefficients. To reduce the dimensionality they used an intraframe linear discriminant analysis and maximum likelihood linear transform (LDA and MLLT), resulting in a 30-dimensional feature vector. To capture dynamic speech information, 15 consecutive feature vectors were concatenated, followed by an interframe LDA/MLLT to obtain dynamic visual features of length 41. They tested their system using the IBM ViaVoice database and reported 17.49% word accuracy in continuous speech recognition. In contrast, Thangthai et al. [220] proposed an ALR system using AAM features and HMM classifiers. Specifically, they trained Context-Independent HMMs (CI-HMM) and Context-Dependent HMMs (CD-HMM), but instead of directly constructing word models, they defined phoneme models. They only report tests on single-speaker experiments in the RM-3000 dataset. A different approach was presented in [32], which used a database with short balanced utterances to define a phoneme-to-viseme mapping able to recognize continuous speech using the VIDTIMIT database. They based their feature extraction on techniques such as PCA or Optical flow, taking into account both the movement and appearance of the lips. On the other hand, [157] used Convolutional Neural Networks (CNNs) to extract high-level features and a combination of HMM to predict phonemes in

spoken Japanese. In yet another work, [113] presented a system based on a set of viseme level HMMs. Concretely, they used Active Appearance Model parameters transformed using LDA as visual features to train their models. They trained 14 HMMs corresponding to 13 visemes plus *Silence* to recover the speech. They tested their method in their own database composed of 1000 words, obtaining 14.08% word accuracy for continuous speech recognition. More recently, [41] collected a very large audio-visual speech database (+100,000 utterances), the Lip Reading Sentences (LRS) database, and proposed a sequence-to-sequence model based solely on CNNs and LSTM networks. They achieved the most significant performance to date in lipreading with 49.8% word accuracy.

We can see that the recognition rates for continuous lip-reading are rather modest in comparison to those achieved for simpler recognition tasks, which can be explained due to the visual ambiguities that appear at the word level. Moreover, continuous lip-reading systems must be able to decode any word of the dictionary and process sentences that contain an arbitrary number of words with unknown time-boundaries, not just pre-defined classes, as is the case when addressing digit-, or word-, or sentence-recognition (at least in the cases in which the targeted classes are a fixed set of predefined phrases).

As mentioned before we are interested in continuous speech recognition because it is the task that is closer to actual lip-reading as done by humans. The available databases for lip-reading in Spanish contain around 600 sentence utterances (+1,000 different words) [164], [61]. Even though results are often not comparable because they are usually reported in different databases, with a variable number of speakers, vocabularies, language and so on, we can consider for comparison to our work, those ALR systems trained with databases with a similar amount of data [183], [220], [157], [113], [171], [228], [210]. However, focusing only on those that address continuous lip-reading (e.g. [113], [183]) we find that word accuracy is typically below the 20%, making evident the big challenges that still remain in this field.

3.1.2 Our System

In this section, each step of our ALR system is explained (Figure 3.1). We start by detecting the face and extracting the region of interest (ROI) that comprises the mouth and its surrounding area. Appearance features are then extracted and used to estimate visemes, which are finally mapped into phonemes with the help of HMMs.

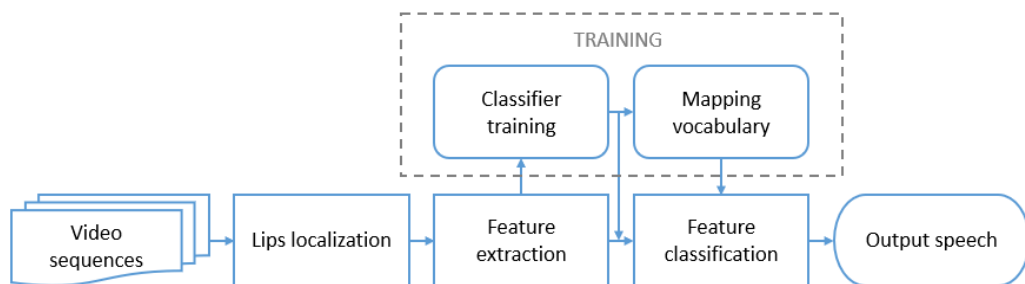


Figure 3.1: General process of an ALR system.

Lips Localization

The location of the face is obtained using invariant optimal features ASM (IOF-ASM) [212] that provides an accurate segmentation of the face in frontal views. The face is tracked at every frame and detected landmarks are used to fix a bounding box around the lips (ROI) (Figure 3.2 (a-b)). At this stage, the ROI can have a different size in each frame. Thus, ROIs are normalized to a fixed size of 48×64 pixels to achieve a uniform representation.

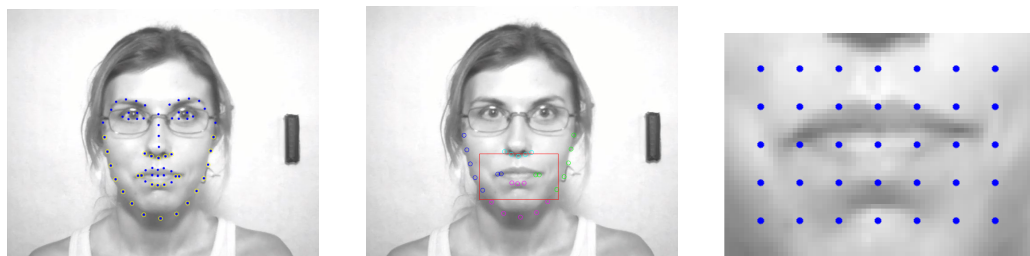


Figure 3.2: (a) IOF-ASM detection, the marks in yellow are used to fix the bounding box; (b) ROI detection, each color fix a lateral of the bounding box; (c) Keypoints distribution.

Feature Extraction

After the ROI is detected a feature extraction stage is performed. Nowadays, there is no universal feature for visual speech representation in contrast to the Mel-Frequency Cepstral Coefficients (MFCC) for acoustic speech. Thus, we look for an informative feature invariant to common video issues, such as noise or illumination changes. We analyze three different appearance-based techniques:

- *SIFT*: SIFT was selected as high level descriptor to extract the features in both the spatial and temporal domains because it is highly distinctive and invariant to image scaling and rotation, and partially invariant to illumination changes and 3D camera viewpoint [124]. In the spatial domain, the SIFT descriptor was applied directly to the ROI, while in the temporal domain it was applied to the centred gradient. SIFT keypoints are distributed uniformly around the ROI (Figure 3.2 (c)). The distance between keypoints was fixed to half of the neighbourhood covered by the descriptor to gain robustness (by overlapping). As the dimension of the final descriptor for both spatial and temporal domains is very high, PCA was applied to reduce the dimensionality of the features. Only statistically significant components (determined by means of Parallel Analysis [68]) were retained.
- *DCT*: The 2D DCT is one of the most popular techniques for feature extraction in ALR [255], [112]. Its ability to compress the relevant information in a few coefficients results in a descriptor with small dimensionality. The 2D DCT was applied directly to the ROI. To fix the number of coefficients, the image error between the original ROI and the reconstructed was used. Based on preliminary experiments, we found that 121 coefficients (corresponding to 1% reconstruction error) for both the spatial and temporal domains produced a satisfactory performance.
- *PCA*: Another popular technique is PCA, also known as *eigenlips* [255], [112], [32]. PCA, similar to 2D DCT is applied directly to the ROI. To decide the optimal number of dimensions the system was trained and tested taking different percentages of the total variance. Lower number of components would lead to a low quality reconstruction, but an excessive number of components will be more affected by noise. In the end 90% of the variance was found to be a good compromise and was used in both spatial and temporal descriptors.

The early fusion of DCT-SIFT and PCA-SIFT has been also explored to obtain a more robust descriptor (see results in Section 3.2.3).

Feature Classification and Interpretation

The final goal of this block is to convert the extracted features into phonemes or, if that is not possible, at least into visemes. To this end we need: 1) classifiers that will map features to (a first estimate of) visemes; 2) a mapping between phonemes and visemes; 3) a model that imposes temporal coherency to the estimated sequences.

1. **Classifiers:** classification of visemes is a challenging task, as it has to deal with issues such as class imbalance and label noise. Several methods have been proposed to deal with these problems, the most common solutions being Bagging and Boosting algorithms [103], [224], [69], [153]. From these, Bagging has been reported to perform better in the presence of training noise and thus it was selected for our experiments. Multiple LDA was evaluated using cross validation. To add robustness to the system, we trained classifiers to produce not just a class label but to estimate also a class probability for each input sample.

For each bagging split, we train a multi-class LDA classifier and use the Mahalanobis distance d to obtain a normalized projection of the data into each class c :

$$d_c(x) = \sqrt{(x - \bar{x}_c)^T \cdot \Sigma_c^{-1} \cdot (x - \bar{x}_c)} \quad (3.1)$$

Then, for each class, we compute two cumulative distributions based on these projections: one for in-class samples $\Phi(\frac{d_c(x) - \mu_c}{\sigma_c})$, $x \in c$ and another one for out-of-class samples $\Phi(\frac{d_c(x) - \mu_{\tilde{c}}}{\sigma_{\tilde{c}}})$, $x \in \tilde{c}$, which we assume Gaussian with means $\mu_c, \mu_{\tilde{c}}$ and variances $\sigma_c, \sigma_{\tilde{c}}$, respectively. An indicative example is provided in Figure 3.3. Notice that these means and variances correspond to the projections in (3.1) and are different from \bar{x}_c and Σ_c .

We compute a class-likelihood as the ratio between the in-class and the out-of-class distributions, as in (3.2) and normalize the results so that the summation over all classes is 1, as in (3.3). When classifying a new sample, we use the cumulative distributions to estimate the probability that the unknown sample belongs to each of the viseme classes (3.3). We assign the class with the highest normalized likelihood L_c .

$$F(c | x) = \frac{1 - \Phi(\frac{d_c(x) - \mu_c}{\sigma_c})}{\Phi(\frac{d_c(x) - \mu_{\tilde{c}}}{\sigma_{\tilde{c}}})} \quad (3.2)$$

$$L_c(x) = \frac{F(c | x)}{\sum_{c=1}^C F(c | x)} \quad (3.3)$$

Once the classifiers are trained we could theoretically try to classify features directly into phonemes, but as explained in Section 3, there are phonemes that share the same visual appearance and are therefore unlikely to be distinguishable by an ALR system. Thus, such phonemes should be grouped into the same class (visemes). In the next subsection we will present a mapping from phonemes to visemes based on the grouping of phonemes that are visually similar.

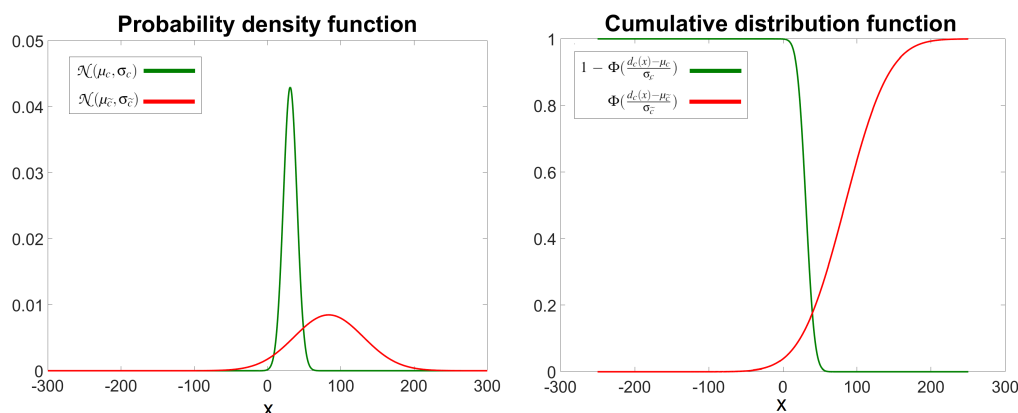


Figure 3.3: (a) Probability density functions for in-class (green) and out-of-class (red) samples; (b) Cumulative distributions corresponding to (a). Notice that for in-class samples we use the complement of the cumulative distribution, since lower values should have higher probabilities. Reprinted from [63].

2. **Phoneme-to-viseme mapping:** to construct our phoneme to viseme mapping we analyze the confusion matrix resulting by comparing the ground truth labels of the training set with the automatic classification obtained from the previous section. We use an iterative process, starting with the same number of visemes as phonemes, merging at each step the visemes that show the highest ambiguity. The method takes into account that vowels cannot be grouped with consonants, because it has been demonstrated that their aggregation produces worse results [32], [18].

The algorithm iterates until the desired alphabet length is achieved. However, there is no accepted standard to fix this value beforehand. Indeed, several different viseme vocabularies have been proposed in the literature typically with lengths between 11 and 15 visemes. Hence, in Section 3.2.3 we will analyse the effect of the alphabet size on recognition accuracy. Once the alphabet construction is concluded, all classifiers are retrained based on the resulting viseme classes.

3. **HMM and Viterbi algorithm:** to improve the performance obtained after feature classification, HMMs of one state per class are used to map: 1) visemes to visemes; 2) visemes to phonemes. An HMM $\lambda = (A, B, \pi)$ is formed by N states and M observations. Matrix A represents the state transition probabilities, matrix B the emission probabilities, and vector π the initial state probabilities. Given a sequence of observation O and the model λ our aim is to find the maximum probability state path $Q = q_1, q_2, \dots, q_{t-1}$.

This can be done recursively using Viterbi algorithm [186], [180]. Let $\delta_i(t)$ be the probability of the most probable state path ending in state i at time t (3.4). Then $\delta_j(t)$ can be computed recursively using (3.5) with initialization (3.6) and termination (3.7).

$$\delta_i(t) = \max_{q_1, \dots, q_{t-1}} P(q_1 \dots q_{t-1} = i, O_1, \dots, O_t | \lambda) \quad (3.4)$$

$$\delta_j(t) = \max_{1 \leq i \leq N} [\delta_i(t-1) \cdot a_{i,j}] \cdot b_j(O_t) \quad (3.5)$$

$$\delta_i(1) = \pi_i \cdot b_i(O_1), 1 \leq i \leq N \quad (3.6)$$

$$P = \max_{1 \leq i \leq N} [\delta_i(T)] \quad (3.7)$$

A shortage of the above is that it only considers a single observation for each time instant t . In our case observations are the output from classifiers and contain uncertainty. We have found that it is useful to consider multiple possible observations for each time step. We do this by adding to the Viterbi algorithm the likelihoods obtained by the classifiers for all classes (e.g. from equation (3.3)). As a result, (3.5) is modified into (3.8), as presented in [63], where the maximization is done across both the N states (as in (3.5)) and also the M possible observations, each weighted with its likelihood estimated by the classifiers.

$$\delta_j(t) = \max_{1 \leq O_t \leq M} \max_{1 \leq i \leq N} [\delta_i(t-1) \cdot a_{i,j}] \cdot \hat{b}_j(O_t) \quad (3.8)$$

$$\hat{b}_j(O_t) = b_j(O_t) \cdot L(O_t) \quad (3.9)$$

The short-form $L(O_t)$ refers to the likelihood $L_{O_t}(x)$ as defined in (3.3). The Viterbi algorithm modified as indicated in (3.8) is used to obtain the final viseme sequence providing at the same time temporal consistency and tolerance to classification uncertainties. Once this has been achieved, visemes are mapped into phonemes using the traditional Viterbi algorithm (3.5). Experimental results of this improvement can be found in [63].

3.2 Experiments

3.2.1 Databases

AV@CAR database

Ortega et al. [164] introduced AV@CAR as a free multi-modal database for automatic audio-visual speech recognition in Spanish, including both studio

Table 3.1: Sample sentences for each database and their corresponding phonetic transcription using SAMPA.

AV@CAR
Francia, Suiza y Hungría ya hicieron causa común. f4'an-Tja sw'i-Ta j uN-g4'i-a jj'a i-Tj'e-4oN k'aw-sa ko-m'un.
Después ya se hizo muy amiga nuestra. des-pw'ez jj'a se 'i-To mw'i a-m'i-Ga nwes-t4a.
Los yernos de ismael no engordarán los pollos con hierba. loz jj'e4-noz De iz-ma-'el n'o eN-go4-Da-4'an los p'o-Los kon jj'e4-Ba.
Me he tomado un café con leche en un bar. me 'e to-m'a-Do 'uN ka-f'e kon l'e-tSe en 'um b'a4.
Guadalajara no está colgada de las rocas. gwa-Da-la-x'a-4a n'o es-t'a kol-G'a-Da De laz r'o-kas.
VLRF
Una sexóloga les ayudó a salvar su relación. 'u-na sek-s'o-lo-Ga les a-jju-D'o a sal-B'a4 su re-la-Tj'on.
Es muy fácil convivir con mis compañeros de piso. 'ez mw'i f'a-Til kom-bi-B'i4 kom mis kom-pa-J'e-4oz De p'i-so.
Cuando tenía quince años fui a mi primer campamento. kwan-do t'e-nja k'in-Te 'a-Jos fw'i a mi p4i-m'e4 kam-pa-m'en-to.
A las ocho de la mañana estaba haciendo pasteles. a las 'o-tSo De la ma-J'a-na es-t'a-Ba a-Tj'en-do pas-t'e-les.
El amanecer es uno de los momentos más bonitos del día. el a-ma-na-T'e4 'es 'u-no De loz mo-m'en-toz m'az Bo-n'i-toz Del d'i-a.

and in-car recordings. The Audio-Visual-Lab dataset of AV@CAR contains sequences of 20 people recorded under controlled conditions while repeating predefined phrases or sentences. There are 197 sequences for each person, recorded in AVI format. The video data has a spatial resolution of 768x576 pixels, 24-bit pixel depth, and 25 fps and is compressed at an approximate rate of 50:1. The sequences are divided into 9 sessions and were captured in a frontal view under different illumination conditions and speech tasks. Session 2 is composed of 25 videos/user with phonetically-balanced sentences. We have used session 2 splitting the dataset into 380 sentences (19 users \times 20 sentences/user) for training and 95 sentences (19 users \times 5 sentences/user) to test the system. Table 3.1 shows 5 samples sentences and their corresponding phonetic transcription.

VLRF database

Fernandez-Lopez et al. [61] introduced VLRF in 2017 as a free multi-speaker database for automatic audio-visual speech recognition in Spanish. The Audio-Visual data contains sequences of 24 people (15 hearing; 9 hearing-impaired) repeating up to three-time sets of 25 sentences selected from a pool of 500 phonetically-balanced sentences (10,000+ word utterances in total). The video data has a spatial resolution of 1280×720 pixels and 50 fps. We have used the first repetition of each sentence per speaker by splitting the dataset into 480 sentences (24 users × 20 sentences/user) for training and 120 sentences (24 users × 5 sentences/user) to test the system. Table 3.1 shows 5 samples sentences and their corresponding phonetic transcription.

3.2.2 Phonetic alphabet

SAMPA is a phonetic alphabet developed in 1989 by an international group of phoneticians, and was applied to European languages as Dutch, English, French, Italian, Spanish, etc. We based our phonetic alphabet in SAMPA because it is the most used standard in phonetic transcription [233], [122]. For the Spanish language, the alphabet is composed by the following phonemes: /p/, /b/, /t/, /d/, /k/, /g/, /tS/, /jj/, /f/, /B/, /T/, /D/, /s/, /x/, /G/, /m/, /n/, /J/, /l/, /L/, /r/, /rr/, /j/, /w/, /a/, /e/, /i/, /o/, /u/. The phonemes /jj/ and /G/ were removed from our experiments because these databases did not contain enough samples to consider them. Table 3.1 shows 10 samples of phonetic transcriptions.

3.2.3 Results

In this section, we show the results of our experiments. In particular, we show the comparison of the performances between the different vocabularies and the different features.

Experimental Setup

We constructed an automatic system that uses local appearance features based on the early fusion of DCT and SIFT descriptors (this combination produced the best results in our tests, see below) to extract the main characteristics of the mouth region in both spatial and temporal domains. The classification of the extracted features into phonemes is done in two steps. Firstly, 100 LDA classifiers are trained using bagging sequences to be robust under label noise. Then, the classifier outputs are used to compute the globally normalized likelihood, as the summation of the normalized likelihood computed by each classifier divided by the number

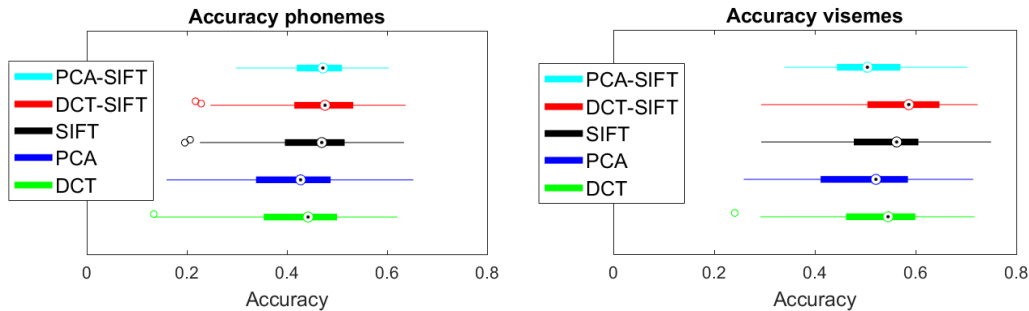


Figure 3.4: Comparison of features performance. Reprinted from [63]

of classifiers (as explained in Section 3.1). Secondly, at the final step, one-state-per-class HMMs are used to model the dynamic relations of the estimated visemes and produce the final phoneme sequences.

Feature Comparison

To analyze the performance of the different features, we extracted DCT, PCA and SIFT descriptors and compared their performance individually and combining DCT-SIFT and DCT-PCA. We used these features as input to 100 LDA classifiers, generated by means of a bagging strategy, and performed a 4-fold cross-validation on the training set. Figure 3.4 displays the results obtained for these experiments on an alphabet of 20 visemes, which was the optimal length in our experiments, as shown in the next section.

Comparing the features independently, DCT and SIFT give the best performances. When combined together, the fusion of both features produced an accuracy of 0.58 for visemes, 0.47 for phonemes.

Comparison of Different Vocabularies

In this section, we investigate the automatic construction of phoneme-to-viseme mappings with the goal to maximize word accuracy. Our system uses these mappings as an intermediate representation which is hypothesized to facilitate the classification of the visual information, given that viseme classes are visually less ambiguous than phoneme classes. At the final step, our system uses HMMs to model the temporal dynamics of the input stream and disambiguate viseme classes based on the sequence context, always producing a final output in terms of phonemes, regardless of the length of the intermediate viseme-based representation.

To evaluate the influence of the different mappings, we analyzed the

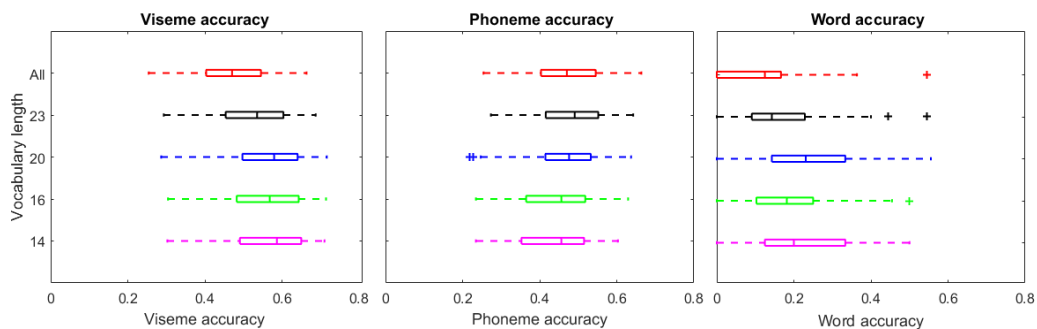


Figure 3.5: Boxplots of system performance the AV@CAR database in terms of viseme-, phoneme- and word accuracy for different vocabularies. We analyze the one-to-one mapping phoneme-to-viseme, and the many-to-one phoneme-to-viseme mappings with 23, 20, 16 and 14 visemes. The phoneme accuracy is always computed from the 28 phonemes.

performance of our system in the AV@CAR database in terms of viseme-, phoneme-, and word-accuracy using viseme vocabularies of different lengths. Our first observation, from Figure 3.5, is that the viseme accuracy tends to grow as we reduce the alphabet length. This is explained by two factors: 1) the reduction in the number of classes, which makes the classification problem a simpler one to solve; 2) the fact that visually indistinguishable units are combined into one. The latter helps to explain the behavior observed in terms of phoneme accuracy. As we reduce the alphabet length, phoneme accuracy firstly increases because we eliminate some of the ambiguities by merging visually similar units. But if we continue to reduce the alphabet, too many phonemes (even unrelated) are mixed together and their accuracy decreases because, even if these visemes are recognized better, their mapping into phonemes is more uncertain. Thus, the optimal performance is obtained for intermediate alphabet lengths, because there is an optimum compromise between the visemes and the phonemes that can be recognized.

A similar effect can be observed in the same figure in terms of words. Firstly, we see that the one-to-one mapping between phoneme and visemes (e.g. using the 28 phonemes classes directly, without merging them into visemes) produces the lowest word accuracy. In contrast, intermediate alphabet lengths show higher word accuracy, with the maximum obtained for 20 classes, supporting the view that the many-to-one mapping from phonemes to visemes is useful to optimize the performance of ALR systems.

Interestingly, while our results support the advantage of combining multiple phonemes into visemes to improve performance, the number of visemes that we

obtain are comparatively high with respect to previous efforts. In our case, the optimal alphabet length for Spanish reduced from 28 phonemes to 20 visemes (including *Silence*), i.e. a reduction rate of about 3 : 2. In contrast, previous efforts reported for English started from 40 to 50 phonemes and merged them into just 11 to 15 visemes [32], which implies reduction rates from 3 : 1 to 5 : 1. It is not clear, however, if the higher compression of the vocabularies obeys to a difference inherent to language or to other technical aspects, such as the ways of defining the phoneme-to-viseme mapping.

Indeed, language differences make it difficult to make a fair comparison of our results with respect to previous work. Firstly, it could be argued that our viseme accuracy is comparable to values reported by [32]; however they used at most 15 visemes while we use 20 visemes and, as shown in Figure 3.5, when the number of visemes decreases, viseme recognition accuracy increases but phoneme accuracy might be reduced, making more difficult to recover the spoken message. Unfortunately, [32] did not report phoneme or word accuracy.

Speaker variability

In the literature, it has been proved that different individuals vocalize in different and unique ways, which results in considerable variability in the difficulty to lip-read across subjects. Thus, it is interesting to compare the performance of the system using different viseme-vocabularies with respect to the different subjects of the database. In Figure 3.6 we show the performance of the phoneme-to-viseme mappings analyzed in the previous section for each of the speakers of the AV@CAR database. We see that, indeed, some speakers are more difficult to lip-read than others, but the relative performance of the different phoneme-to-viseme mappings varies only marginally. Specifically, it can be observed that the 20-visemes alphabet obtains the highest word accuracy for the majority of speakers in the database.

3.3 Discussion

Visual ambiguities have been one of the most investigated problems in ALR. In Section 3, we described the minimum auditory units (phonemes) and their visual equivalent (visemes), as well as their many-to-one relation. Focusing on visemes, there exist two different points of view in the literature: i) researchers that defend their utility and proposed several phoneme-to-viseme mappings [32], [63], [66], [89], [18]; ii) researchers that debate their actual usefulness and existence [41], [194], [42], [11].

In this work, we proposed to automatically construct a phoneme-to-viseme

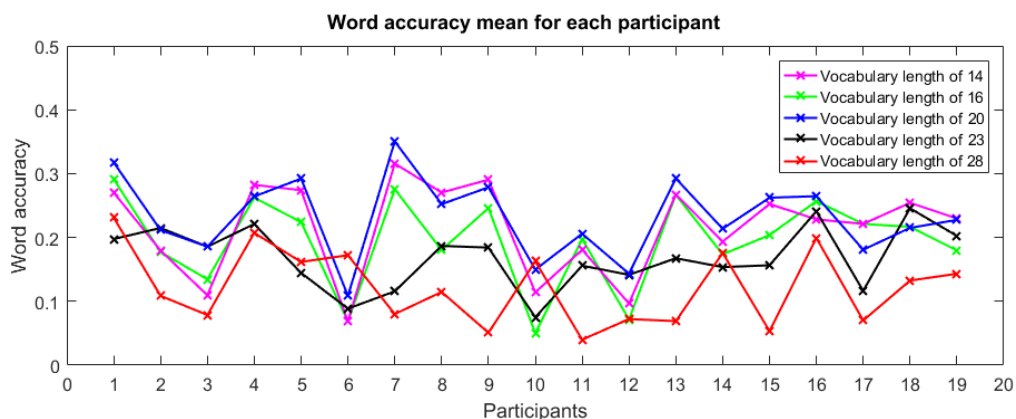


Figure 3.6: Comparison of system performance in the AV@CAR database in terms of word accuracy for the different vocabularies and participants.

mapping based on visual similarities between phonemes to maximize word accuracy. Thus, we investigated the usefulness of different phoneme-to-viseme mappings, obtaining the best results for intermediate alphabet lengths. However, it is also interesting to analyze additional qualitative and quantitative results, such as lip crops and confusion matrices.

Firstly, we can find intuitive support to the existence of visemes by visually analyzing the lips of subjects when pronouncing different phonemes. In Table 3.2 we show examples of the average lip-images per subject and phoneme for 5 subjects from the AV@CAR database. That is, each cell of the table contains the average of all frames for which a given subject was uttering a certain phoneme. Looking at the examples, we can clearly see strong visual similarities between some of the phonemes. For example, it would be arguably difficult to distinguish between the averages from /a/ and /e/, or between /m/, /p/ and /B/, which correlates well with the proposed viseme mappings. On the other hand, even when there exist visual similarities between /o/ and /u/, we can observe that for /u/ there appears to be a smaller hole inside the lips than for /o/ in most of the cases. The latter suggests that these two phonemes might actually be visually separable, but in our experiments the classification results showed considerable confusion between them (see also Figure 3.7) and the best performance was obtained with an alphabet in which /o/ and /u/ were merged into the same viseme.

Another interesting observation from Table 3.2 is the variability between subjects. For example, in the first two subjects in the table, the averages for the phoneme /tS/ seem slightly different from the averages for /t/ and /s/; while the other subjects show extremely similar averages, that are arguably indistinguishable. This observation is in line with the discussion from the previous

Table 3.2: Average lip-images per user and phoneme of 5 subjects of the AV@CAR database. Each row shows a sample subject. For each subject, every column shows the average of all the frames in which the subject uttered a specific phoneme. The vertical lines separate the phonemes that belong to different visemes according to our mapping. The last row shows the average when considering all the users together.





























































a	e	m	p	B	o	u	t	s	tS
Average across subjects									

section, i.e. the fact that each person vocalizes in a unique way and there are subjects that are easier to lip-read than others.

Thus, it is interesting to compare the preceding lip-images with those recorded by people who consciously try to vocalize well to be easily lip-read. For this purpose, we decided to analyze also the lip-images from the Visual Lip Reading Feasibility (VLRf) database [61]. Similarly to AV@CAR, the VLRf database is an audiovisual database recorded in Spanish in which speakers were recorded while reading a series of sentences that were provided to them. However, while in AV@CAR subjects were speaking naturally, in the VLRf database speakers were instructed to make their best effort to be easily understood by lip-reading. Hence, we could hypothesize that, if it were true that all phonemes are visually distinguishable (which would imply that there is no need for visemes) then the VLRf would be an ideal corpus to visualize this.

To test the above hypothesis, we replicated our experiments in the VLRf

Table 3.3: Average lip-images per user and phoneme of 5 subjects of the VLRf database. Each row shows a sample subject. For each subject, every column shows the average of all the frames in which the subject uttered a specific phoneme. The vertical lines separate the phonemes that belong to different visemes according to our mapping. The last row shows the average when considering all the users together.

a	e	m	p	B	o	u	t	s	tS
									
									
									
									
									
Average across subjects									
									

database to make them directly comparable to those from the AV@CAR database. We start by showing the obtained results in Table 3.4 while Table 3.3 shows examples of the average lip-images per subject and phoneme for 5 subjects from the VLRf database. Compared to those in Table 3.2, we still observe the same visually similar units, that correlate with our mappings. However, looking separately at each speaker (e.g. each row of the table), we also observe that some of the phonemes seem now more likely to be distinguished. For example, even though phonemes /a/ and /e/ produce very similar lip-images, in /a/ the mouth seems more open vertically while in /e/ the mouth seems widened (more horizontal opening). It is also possible to find differences between /t/, /s/ and /tS/, e.g. /t/ seems to be more open with visibility of the tongue and /tS/ seems to be pronounced joining the lips more strongly. However, this is not true for all phonemes, e.g. the differences between /m/, /p/ and /B/ are still visually imperceptible. Moreover, the differences between phonemes from the same

Table 3.4: System performance in the VLRF database in terms of viseme-, phoneme- and word accuracy for the vocabularies of 20 and 28 classes.

Alphabet length	Viseme accuracy	Phoneme accuracy	Word accuracy
20	56.07%	51.25%	20.76%
28	51.78%	51.78%	18.17%

speaker do not necessarily generalize across multiple speakers, as we can see in the last row of Table 3.3: the lip-images averaged across multiple speakers are again extremely similar, reflecting the visual ambiguities that justify the mapping of groups of phonemes into the same viseme. As a result, even in a dataset in which subjects were trying to vocalize clearly to facilitate lip-reading, the visual ambiguities between phonemes are still very difficult to distinguish and additional information related to the context would be required to disambiguate them.

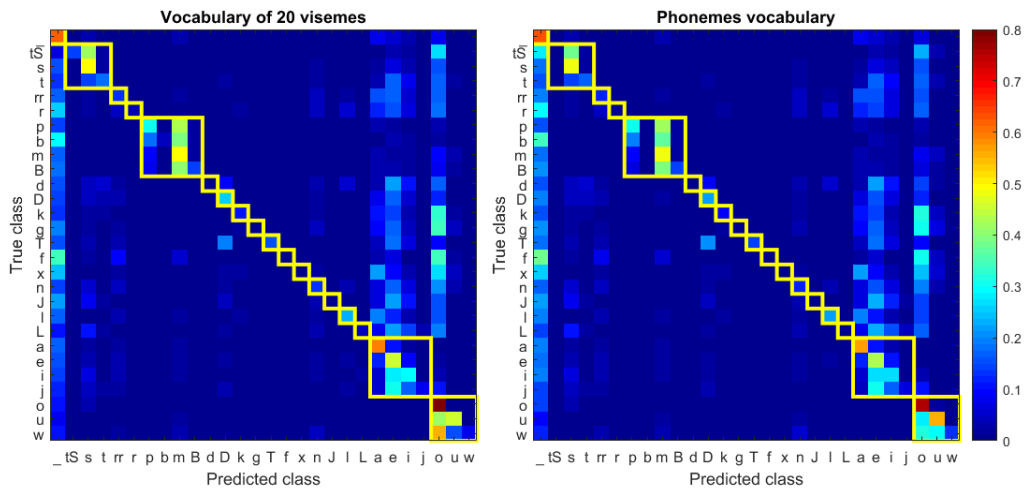


Figure 3.7: (a) Resulting confusion matrix from a system trained in VLRF using 20 visemes (many-to-one phoneme-to-viseme mapping). (b) Resulting confusion matrix from a system trained in VLRF using 28 visemes (one-to-one phoneme-to-viseme mapping). Additionally, we highlighted in yellow, the phonemes that share the same viseme in the proposed alphabet to a clearer comprehension.

A similar conclusion is achieved when we analyze the results in quantitative terms, by looking at the confusion matrices obtained with and without using visemes as an intermediate representation. Specifically, Figure 3.7 shows the confusion matrices of our ALR system trained in two ways: firstly, using a phoneme-to-viseme mapping of 20 visemes (found to be optimal experimentally), and the second one trained using a one-to-one mapping between phonemes and

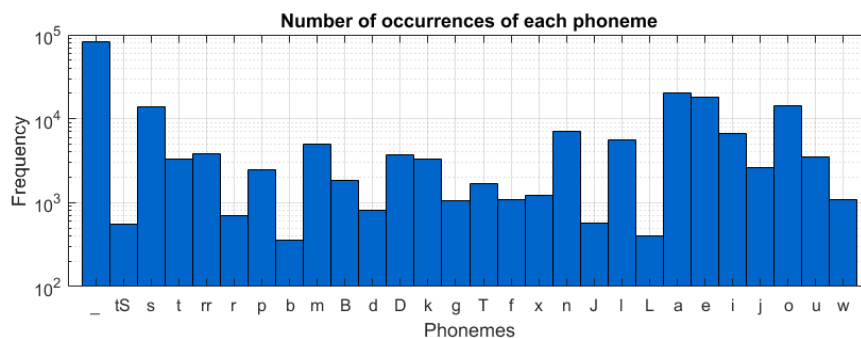


Figure 3.8: Frequency of appearance of each phoneme in the VLRf database.

visemes (i.e. without visemes). Notice, however, that in both cases we evaluate the confusion at the final stage of the system, which always produces phoneme estimates. Thus, both confusion matrices show the performance of the system in terms of phonemes. A notable observation from these two matrices is that there is high confusion between phonemes that map into the same viseme. This behaviour would be expected in the first matrix, as it corresponds to a system trained based on such phoneme-to-viseme mappings. However, we also see that a very similar confusion appears also in the second matrix, even when the system was trained directly on phonemes in all its stages.

Detailed analysis of Figure 3.7 highlights a few other interesting points. Firstly, in some cases the confusion between groups of phonemes are not symmetric, e.g. although the phonemes /s/ and /t/ are visually similar, the system outputs /s/ more often than /t/, probably because the first one has a higher frequency of appearance in the training set (see Figure 3.8). Secondly, there is a huge confusion between several consonants that are very often misclassified as vowels by the system. This type of confusion does not seem directly related to visual similarities, but to difficulties in labeling phoneme transitions and to class imbalance. On the one hand, it is very difficult to precisely define the boundaries between consecutive phonemes and, additionally, these can be influenced by previous and posterior phonemes, which leads to ambiguous labelling. The considerably higher number of vowel samples when compared to consonants explains why the confusion is not symmetric and vowels are rarely misclassified as consonants, except for phonemes with a comparably high number of samples, e.g. /s/, /m/, /n/.

3.4 Conclusions

We investigate the automatic construction of optimal viseme vocabularies by iteratively combining phonemes with similar visual appearance into visemes. We perform tests on the Spanish databases AV@CAR and VLRF using an ALR system based on the combination of DCT and SIFT descriptors and HMMs to model both viseme and phoneme dynamics. Using 19 and 24 different speakers, respectively for AV@CAR and VLRF, we reach a 58% and 56% of recognition accuracy in terms of viseme units, 47% and 51% in terms of phoneme units and 23% and 21% in terms of words units.

Our experiments support the advantage of merging groups of phonemes into visemes. We find that this is the case of both for phonemes that are visually indistinguishable (e.g. /b/, /m/ and /p/) as well as for those in which it is possible to perceive subtle but insufficient differences. The latter occurs, for example, in the case of the phonemes /s/, /t/ and /tS/, for which it is possible to identify visual differences within the same subject but these do not seem to reproduce consistently across multiple subjects. Moreover, experiments on a comparable ALR system trained exclusively using phonemes at all its stages confirmed the existence of strong visual ambiguities between groups of phonemes. This fact and the higher word accuracy obtained when using phoneme-to-viseme mappings, justify the usefulness of visemes instead of the direct use phonemes.

Thus, even though going through visemes may seem like a loss of information, this is only partially true because looking at independent time instants (or small-time windows) there is no perceivable difference, in visual terms, between some phonemes. Therefore, training a classifier to predict phonemes based on such information seems like an ill-posed problem, since mistakes between arguably non-separable classes (phonemes within the same viseme) contribute to the loss function as much as those from separable ones (different visemes). Once we estimate the viseme classes, we can disambiguate them into phonemes by means of word or sentence context (e.g. by using HMMs or, more recently, Recurrent Neural Networks).

Chapter 4

THE UPPER BOUND OF VISUAL SPEECH RECOGNITION

Speech is the most used communication method between humans, and it is considered a multi-sensory process that involves the perception of both acoustic and visual cues since McGurk demonstrated the influence of vision in speech perception. Many authors have subsequently demonstrated that the incorporation of visual information into speech recognition systems improves their robustness [143, 183].

Visual information usually involves position and movement of the visible articulators (the lips, the teeth and the tongue), speaker localization, articulation place and other signals not directly related to the speech (facial expression, head pose and body gestures) [89, 236, 37]. Even though the audio is in general much more informative than the video signal, speech perception relies on the visual information to help decoding spoken words as auditory conditions are degraded [89, 54, 213, 191]. Furthermore, for people with hearing impairments, the visual channel is the only source of information to understand spoken words if there is no sign language interpreter [183, 10, 200]. Therefore, visual speech recognition is implicated in our speech perception process and is not only influenced by lip position and movement but it also depends on the speaker's face, as it has been shown that it can also transmit relevant information about the spoken message [236, 37]. Much of the research in Automatic Speech Recognition (ASR) systems have focused on audio speech recognition, or on the combination of both modalities using Audio-Visual Speech Recognition (AVSR) systems to

Adapted from: Fernandez-Lopez, A., Martinez, O., & Sukno, F. M. (2017, May). Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database. *In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 208-215). IEEE. DOI: 10.1109/FG.2017.34.

improve the recognition rates, but Visual Speech Recognition (VSR) systems have been less frequently analyzed alone [51, 151, 255, 246, 210, 41, 173, 7]. The performance of audio-only ASR systems is very high if there is not much noise to degrade the signal. However, in noisy environments AVSR systems improves the recognition performance when compared to their audio-only equivalents [183, 51]. In contrast, in visual-only ASR systems the recognition rates are rather low [254]. This can be partially explained by the higher difficulty associated with decoding speech through the visual channel, when compared to the audio channel.

One of the key limitations of VSR systems resides in the ambiguities that arise when trying to map visual information into the basic phonetic unit (phonemes), i.e. not all the phonemes that are heard can be distinguished by observing the lips. There are two types of ambiguities: *i*) there are phonemes that are easily confused because they look visually similar between them (e.g: /p/, /b/ and /m/). For example, the phones /p/ and /b/ are visually indistinguishable because voicing occurs at the glottis, which is not visible; *ii*) there are phonemes whose visual appearance can change (or even disappear) depending on the context. This is the case of the *velars*, consonants articulated with the back part of the tongue against the soft palate (e.g: /k/ or /g/), because they change their position in the palate depending on the previous or following phoneme. Specifically, velar consonants tolerate palatalization (the phoneme changes to palatal) when the previous or following phoneme is a vowel or a palatal [146]. Other drawbacks associated with lipreading have also been reported in the literature, such as the distance between the speakers, illumination conditions or visibility of the mouth [89, 29, 165]. However, the latter can be easily controlled, while the ambiguities explained above are limitations intrinsic to lip-reading and constitute an open problem.

On the other hand, it is known that some people are very good lip-readers. In general, visual information is the only source of reception and comprehension of oral speech for people with hearing impairments, which leads to the common misconception that they must be good lip-readers. Indeed, while many authors have found evidence that people with hearing impairments outperform normal-hearing people in comprehending visual speech [182, 21, 31, 52, 133], there are also several studies where no differences were found in speech-reading performance between normal-hearing and hearing-impaired people [190, 110]. Such conflicting conclusions might be partially explained by the influence of other factors beyond hearing impairment. For example, it is well known that human lip-readers use the context of the conversation to decode the spoken information [89, 37, 29], thus it has been argued that people who are good lip-readers might be more intelligent, with more knowledge of the language, and with a more comprehensible oral speech for others [165, 190, 145, 111].

While the above complexities may provide some explanation to the rather low recognition rates of VSR systems, there seems to be a significant gap between

these and human lip-reading abilities. More importantly, it is not clear what would be the upper bound of visual-speech recognition, especially for systems not using context information (it has been argued that humans can *read* only around 30% of the information from the lips, and the rest is filled-in from the context [165, 50]). Thus, it is not clear if the poor recognition rates of VSR systems are due to inappropriate or incomplete design or because there is an intrinsic limitation in visual information that causes the impossibility of perfect decoding of the spoken message.

Contributions: In this work we explore the feasibility of visual speech reading with the aim to estimate the recognition rates achievable by human observers under favorable conditions and compare them with those achieved by an automatic system. To this end, we focus on the design and acquisition of an appropriate database in which recorded speakers actively aim to facilitate lip-reading but conversation context is minimized. Specifically, we present a new database recorded with the explicit goal of being visually informative of the spoken message. Thus, data acquisition is especially designed with the aim that a human observer (or a system) can decode the message without the help of the audio signal. Concretely, lip-reading is applied to people that are aware of being read and have been instructed to make every effort so that they can be understood based exclusively on visual information. Then, the database deals with sentences that are uttered slowly, with repetitions, well pronounced and viewed under optimal conditions ensuring good illumination and mouth visibility (without occlusions and distractions).

In this database we divided the participants into two groups: 9 hearing-impaired subjects and 15 normal-hearing subjects. In our tests, hearing-impaired participants outperformed the normal-hearing participants but without reaching statistical significance. Human observers outperform markedly the VSR system in terms of word recognition rates, but in terms of phonemes, the automatic system achieves very similar accuracy to human observers.

4.1 Audio-visual speech databases

Visual only speech recognition spans over more than thirty years, but even today is still an open problem in science. One of the limitations for the analysis of VSR systems is the accessible data corpora. Despite the abundance of audio speech databases, there exist a limited number of databases for audio-visual or visual only ASR research. That is explained in the literature because the field is relatively young, and also, because the audio-visual databases add some challenges such as database collection, storage and distribution, not found as a problem in audio corpora. Acquisition of visual data at high resolution, frame rate and image

quality, with optimal conditions and synchronized with the audio signal requires expensive equipment. In addition, visual storage is at least one or two orders of magnitude to the audio signal, making his distribution more difficult [255], [184].

Most databases used in audio-visual ASR systems suffer from one or more weaknesses. For example, they contain low number of subjects ([140, 47]), small duration ([140, 47, 116, 144]), and are addressed to specific and simple recognition tasks. For instance, most corpora are centered in simple tasks such as isolated or connected letters ([140, 47, 116]), digits ([116, 144, 170, 96, 128]), short sentences ([144, 198, 46, 142, 250, 9]) and only recently continuous speech ([96, 164, 87, 25]). These restrictions make more difficult the generalization of methods and the construction of robust models because of the few samples of training. Additional difficulties are that some databases are not freely available.

As explained in Section 4 the aim of this project is to apply continuous lip-reading to people that are conscious of being read and is trying to be understood based exclusively on visual information. Thus, from the most common databases, only VIDTIMIT [198], AVICAR [116], Grid [46], MOBIO [142], OuluVS [250], OuluVS2 [9], AV@CAR [164], AV-TIMIT [87], LILiR [25] contain short sentences or continuous speech and could be useful to us. However, we rejected the use of them because the participants speak in normal conditions without previous knowledge of being lip-read. In addition, most of the databases have low technical aspects and a limited number of subjects with restricted vocabularies centred in repetitions of short utterances. Subsequently, we decided to develop a new database designed specifically for recognizing continuous speech in controlled conditions.

4.2 Visual Lip-Reading Feasibility Database

The Visual Lip-Reading Feasibility (VLRf) database is designed with the aim to contribute to research in visual-only speech recognition. A key difference of the VLRf database with respect to existing corpora is that it has been designed from a novel point of view: instead of trying to lip-read from people who are speaking naturally (normal speed, normal intonation,...), we propose to lip-read from people who strive to be understood.

Therefore, the design objective was to create a public database visually informative of the spoken message in which it is possible to directly compare human and automatic lip-reading performance. For this purpose, in each recording session there were two participants: one speaker and one lip-reader. The speaker was recorded by a camera while pronouncing a series of sentences that were provided to him/her; the lip-reader was located in a separate room, acoustically isolated from the room where the speaker was located. To make the human

decoding as close as possible to the automatic decoding, the input to the lip-reader was exclusively the video stream recorded by the camera, which was displayed in real-time by means of a 23" TV screen.

After each uttered sentence, the lip-reader gave feedback to the speaker (this was possible because it was possible to enable audio feedback from the lip-reading room to the recording room, but not conversely). Each sentence could be repeated up to 3 times, unless the lip-reader decoded it correctly in fewer repetitions. Both the speaker utterances and the lip-reader answers (at each repetition) were annotated.

Participants were informed about the objective of the project and the database. They were also instructed to make their best effort to be easily understood, but using their own criteria (e.g: speak naturally or slowly, emphasize the separation between words, exaggerate vocalization,...).

Each recording session was divided into 4 levels of increasing difficulty: 3 levels with 6 sentences and 1 level with 7 sentences. We decided to divide the session in different levels to make it easier for participants to get accustomed to the lip-reading task (and perhaps also to the speaker). Specifically, in the first level the sentences are short with only a few words, and as the level increases the difficulty increases in terms of number of words. The sentences are unrelated among them and only the context within the sentence is present. Thus, in the first sentences participants had to read fewer words but with very little context and in the last sentences the context was considerably more important and would certainly help decoding the sentence. To motivate participants and to ensure their concentration during all the session, at the end of each level both participants changed their roles.

Finally, because our objective was to determine the visual speech recognition rates that could be achievable, we also recruited volunteers which were hearing-impaired and accustomed to use lip-reading in their daily routine. Then, we will also compare the capability of lip-reading of normal-hearing and hearing-impaired people.

4.2.1 Participants

We recruited 24 adult volunteers (3 male and 21 female). Thirteen are University students, one is Teacher of Sign Language at UPF and the other 10 participants are members of the Catalan Federation of Associations of Parents and Deaf (ACCAPS) [1]. The 24 participants were divided in two groups: normal-hearing people and hearing-impaired people.

– *Normal-hearing participants.* Fifteen of the volunteers are normal-hearing participants (14 females and 1 male), who were selected from a similar educational range (e.g: same degree) because, as explained in Section 4, lip-

reading abilities have been related to intelligence and language knowledge. Two of the participants were more than 50 years old and have a different education level while the other 13 subjects of this group shared educational level and age range.

– *Hearing-impaired participants.* There were nine hearing-impaired participants, all above 30 years old (7 female and 2 male). Eight of them have post-lingual deafness (the person loses hearing after acquiring spoken language) and one has pre-lingual deafness (the person loses hearing before the acquisition of spoken language). There were 4 participants with cochlear implants or hearing aids.

4.2.2 Utterances

Each participant was asked to read 25 different sentences, from a total pool of 500 sentences, proceeding similarly to [46]. The sentences were unrelated between them to avoid that lip-readers could benefit from conversation context. Sentences had different levels of difficulty, in terms of their number of words. There were 4 different levels, from 3-4 words, 5-6 words, 7-8 words and 8-12 words. We decided to divide the sentences into different levels for two reasons. Firstly, to allow lip-readers to get some *training* with the short sentences of the first level (i.e. to get acquainted and gain confidence with the setup, the task and the speaker). Secondly, to compare the effect of the context in the performance of human lip-readers. The utterances with fewer words have very little context, while longer sentences contained considerable context that should help the lip-reader when decoding the message.

Overall, there were 10200 words in total (1374 unique), with an average duration of 7 seconds per sentence and a total database duration of 180 minutes (540,162 frames). The sentences contained a balanced phonological distribution of the Spanish language, based on the balanced utterances used in the AV@CAR database [164].

4.2.3 Technical aspects

The database was recorded in two contiguous soundproof rooms (Fig. 4.1). The distribution of the recording equipment into the rooms is shown in Fig. 4.1. A Panasonic HPX 171 camera was located with a tripod PRO6-HDV in front of the chair of the speaker, to ensure an approximately frontal face shot, with a supplementary directional microphone mounted on the camera to ensure a directional coverage in the direction of the speaker. The camera recorded a close up shot (Fig.4.1) at 50 fps with a resolution of 1280×720 pixels and audio at



Figure 4.1: Scheme of the recording setup and snapshots of the VLRF database.

48 kHz mono with 16-bit resolution. Two Lumatek ultralight 1000W Model 53-11 were used together with reflecting panels to obtain a uniform illumination and minimize shadows or other artifacts on the speaker's face. When performing the lip-reading task, the lip-reader was located in the control room. The position of the lip-reader was just in front of a 23" LG Flatron M2362D PZ TV. This screen was connected to the camera so that it reproduced in real-time what the camera was recording. Only the visual channel of the camera was fed into the control room, although both audio and video channels are recorded for post-processing of the database. The rooms were acoustically isolated between them except for the feedback channel composed by a microphone in the control room and a loudspeaker in the recording room. This channel was used after each utterance to let the speaker know what message was decoded by the lip-reader.

4.2.4 Data labeling

The ground-truth of the VLRF database consists of a phoneme label per frame. We used the EasyAlign plug-in from Praat [24], which allows to locate the phoneme in each time instant based on the audio stream. Specifically, the program locates the phonemes semi-automatically and there is usually the need for manual intervention to adapt the boundaries of each phoneme to more precise positions. The phonemes used are based on the phonetic alphabet SAMPA [233]. For the Spanish language, the SAMPA alphabet is composed of the following 31 phonemes: /p/, /b/, /t/, /d/, /k/, /g/, /tS/, /j/, /f/, /B/, /T/, /D/, /s/, /z/, /x/, /G/, /m/, /n/, /N/, /J/, /l/, /L/, /t/, /4/, /j/, /w/, /a/, /e/, /i/, /o/, /u/.

4.3 Results

In this section we show the word- and phoneme-recognition rates obtained in our experiments. We start by analyzing the human lip-reading abilities and comparing the performance of hearing-impaired and normal-hearing participants. Then, we analyse the influence of training and context in human performance. Finally, we compare the performance of our automatic system to the results obtained by human observers.

The use of two separate measures (word and phoneme rates) is necessary to analyze different aspects of our results. On one hand, phonemes are the minimum distinguishable units of speech and directly constitute the output of our automatic system. However, the ultimate goal of lip-reading is to *understand* the spoken language, hence the need to focus (at least) on words. It is important to notice that acceptable phoneme recognition rates do not necessarily imply good word recognition rates, as will be shown later.

The word recognition rate was computed as the fraction of words correctly understood in a given sentence. The phoneme recognition rate was computed as the fraction of video frames in which the correct phoneme was assigned. Consequently, 25 accuracy measures were computed for each participant and each repetition. Recognition rates for the automatic system were computed in the same manner, except that there were no multiple repetitions.

4.3.1 Experimental setup

Our VSR system starts by detecting the face and performing an automatic location of the facial geometry (landmark location) using the Supervised Descend Method (SDM) [241]. Once the face is located, the estimated landmarks are used to fix a bounding box around the region (ROI) that is then normalized to a fixed size. Later on, local appearance features are extracted from the ROI based on early fusion of DCT and SIFT descriptors in both spatial and temporal domains. As explained in Section 4 there are phonemes that share the same visual appearance and should belong to the same class (visemes). Thus, we constructed a phoneme to viseme mapping that groups 32 phonemes into 20 visemes based on an iterative process that computes the confusion matrix and merges at each step the phonemes that show the highest ambiguity until the desired length is achieved. Then, the classification of the extracted features into phonemes is done in two steps. Firstly, multiple LDA classifiers are trained to convert the extracted features into visemes and secondly, at the final step, one-state-per-class HMMs are used to model the dynamic relations of the estimated visemes and produce the final phoneme sequences. This system was shown to produce near state-of-the-art performance

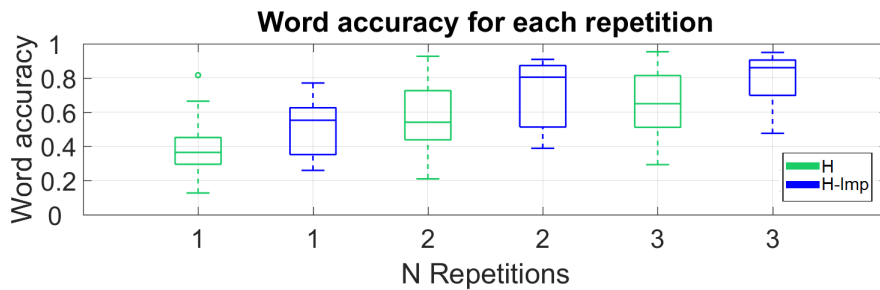


Figure 4.2: Word accuracy for normal-hearing (H) and hearing-impaired groups (H-Imp) at each repetition.

for continuous visual speech-reading tasks (more details in [64]).

4.3.2 Human lip-reading

As explained in Section 4, it is not clear if hearing-impaired people are better lip-readers than normal-hearing people. Fig. 4.2 shows the word recognition rates for both groups at each repetition and Fig. 4.3 shows the word recognition rates for each participant and repetition. Analyzing each participant individually, it is difficult to observe any group-differences between hearing-impaired and normal-hearing participants. However, we do observe large performance variations within each of the groups, i.e. there are very good and quite poor lip-readers regardless of their hearing condition.

On the other hand, looking at the results globally, split only by group (Fig. 4.2), they suggest that hearing-impaired participants outperform normal-hearing participants in the lip-reading task for all three repetitions. However, the results differ about 20% in terms of word recognition rate and thus we need to study if this difference is statistically significant.

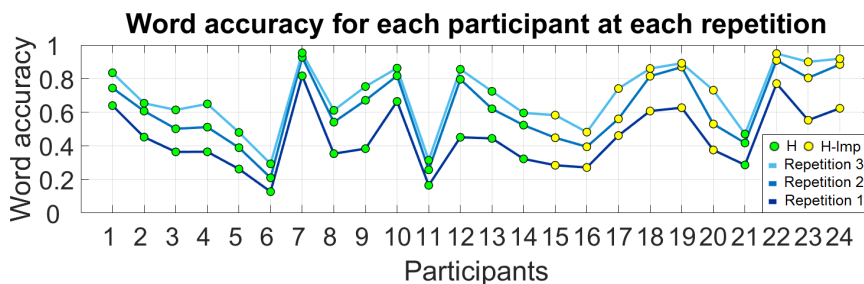


Figure 4.3: Word accuracy per participant at each repetition.

To do so, we estimated the word accuracy of each participant as the average

Table 4.1: Statistical comparison between hearing-impaired and normal-hearing participants at each repetition.

Attempt	Wilcoxon signed rank	Unpaired two-sample
1	p = 0.116	p = 0.094
2	p = 0.094	p = 0.088
3	p = 0.041	p = 0.037

accuracy across the 25 sentences that he/she had to lip-read. Then, we performed statistical tests to determine if there were significant differences between the 9 hearing-impaired samples and the 15 normal-hearing samples. Because we only want to test if the hearing-impaired participants were better than normal-hearing participants, we performed single-tailed tests where the null hypothesis was that the mean or median (depending on the test) performance of hearing-impaired participants was not higher than the performance of normal-hearing participants. We ran two tests (summarized in Table 4.1) for each of the 3 repetitions: *Wilcoxon signed rank test* and *Unpaired two-sample t-test*. Taking the conventional significance threshold of $p < 0.05$ it could be argued that at the third repetition the performance of hearing-impaired participants was significantly better than that of normal-hearing participants. However, this was not observed in the first two repetitions. Moreover, the 9 hearing-impaired subjects did better than the 15 normal-hearing, but taking into account that the sample size is relatively small, current trends in statistical analysis suggest that the obtained p-values are not small enough to claim that this would extrapolate to the general population. On the other hand, looking at the p-values, with the current number of subjects we are not far from reaching significance [45].

In Figures 4.2 and 4.3 we also show the influence of repetitions into the final performance: as the number of repetitions increases the recognition rate increases too. This effect can be seen split by group and analysing each participant separately.

4.3.3 Training and context influence on lip-reading

The context is one of the human resources more used in lip-reading to complete the spoken message. To analyse the influence of the context, the participants were asked to read four different types of sentences, in terms of number of words (explained in Section 4.2). Thus, as the level increases, sentences are longer and the context increases too.

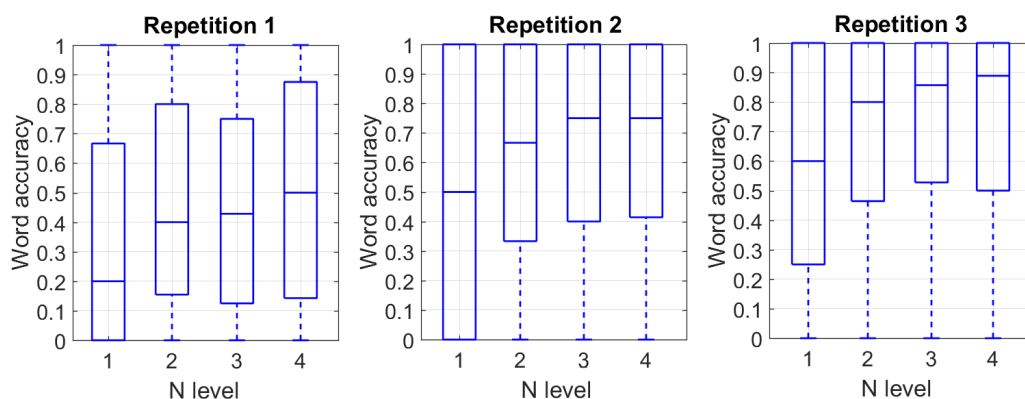


Figure 4.4: Word recognition average for each participant at each level.

In Fig. 4.4 we can observe how the first level has the lowest word recognition rates for all repetitions, while the last level has the highest rates. There are two factors that could contribute to this effect: 1) Context: humans use the relation between words to try decoding a meaningful message, and 2) Training: as the level increases the participants are more acquainted to the speaker and to the lip-reading task.

The results of Fig. 4.4 are not enough to determine whether the effect is due to context, training or both. Thus, in Fig. 4.5 we analyze the variation of performance per sentence (with a cumulative average) instead of per level, which should make clearer the effect of training. This is because training occurs continuously from one sentence to another while context only increases when we change from one level to the next one. Thus, the effect of training can be seen as the constant increase performance in each of the curves (up to 20%). As the users have lip-read more sentences they tend to become better lip-readers. On the other hand, the influence of context is better observed by comparing the different repetitions. In the first attempt, the sentence was completely unknown to the participants, but, in the second and third repetitions there was usually some context available because the message had been already partially decoded, hence constraining the possible words to complete the sentence.

4.3.4 Human observers and automatic system comparison

The results of the automatic system are only computed for the first attempt, since it was not designed to benefit from repetitions. The resulting word-recognition rates are shown in Fig. 4.6 (Top). Notice that now the participant number indicates the person that was pronouncing the sentences as the recognition is always performed by the system. Thus, this figure provides information about how well the system

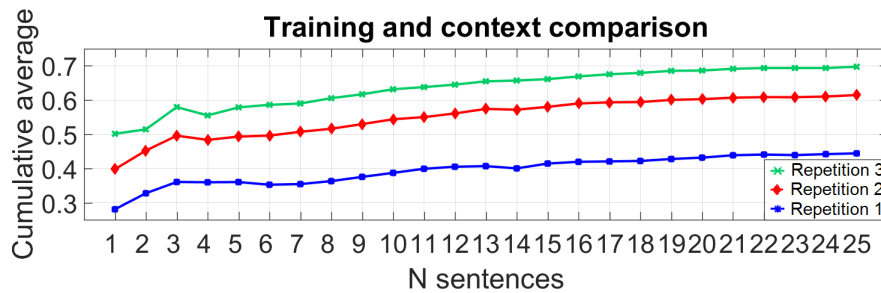


Figure 4.5: Cumulative average per sentence for all participants at each repetition.

was able to lip-read each of the participants. The system produced the highest recognition rates for participants 1, 8, 17 and 21. Interestingly, these participants had good pronunciation and visibility of the tongue and teeth.

We are interested in comparing the performance of humans lip-reading and a VSR system. Focusing on Fig. 4.7 (Top) we can observe how the word recognition rates are lower for the system in most of the cases. However, we have to take into account that the system does not use the context into the sentence. Indeed, the system is not even targeting words but phonemes, which are later merged to form words. In contrast, people directly search for correlated words with the lip movements of the speaker. Thus, it is reasonable to expect a considerable gap between human and automatic performance, which will be shown to reduce considerably if the comparison is done in terms of phonemes.

In the same figure (Fig. 4.7) we can observe a direct comparison of the mean recognition rates of each participant identified by humans and by the automatic system. The system gives an unbiased measure of the facility to lip-read participants because it evaluates each of them in the same manner. In contrast, human lip-reading was performed in couples (couples are organized in successive order, e.g. participants 1 and 2, 3 and 4, etc), hence each participant was only lip-read by its corresponding partner. Analyzing Fig. 4.7 we can identify which users were good lip-readers and also good speakers. For example, participant 7 was lip-read by participant 8 with a high word recognition rate. Then, in the curve corresponding to human performance, we observe a high value for participant 8, meaning that he/she was very successful at lip-reading. When we look at the system's performance, however, the value assigned to participant 8 corresponds to the rate obtained by the system and is therefore a measure related to how participant 8 spoke rather than how he/she lip-read. For this specific participant, the figure shows that system performance was also high, hence he/she is a candidate to be a good lip-reader and speaker.

The word recognition rates reported by our system are rather low compared to those obtained by human observers. However, as stated earlier, our system is

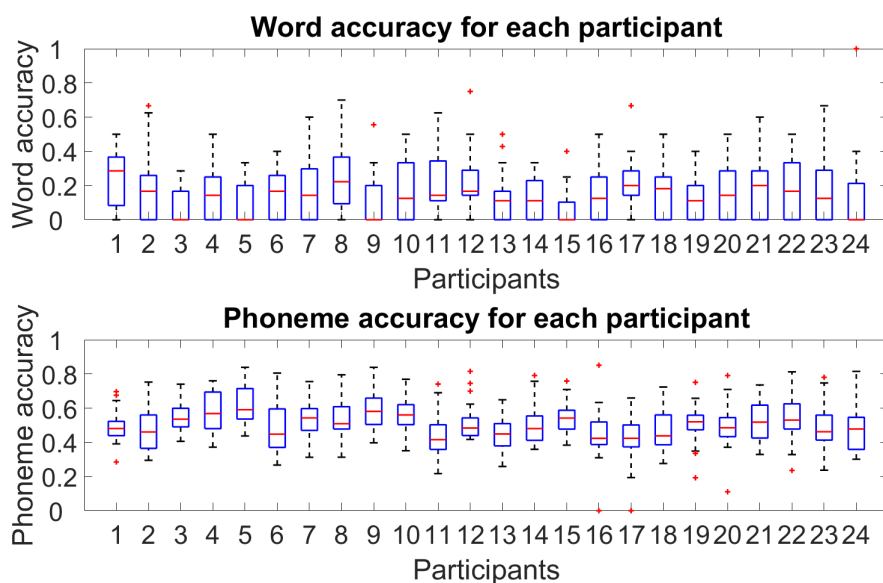


Figure 4.6: Top: system performance in terms of word recognition rate for each participant. Bottom: system performance in terms of phoneme recognition rate for each participant.

trying to recognize phonemes and convert them to words, so it is also interesting to analyze its performance in terms of phoneme recognition. The phoneme recognition rates obtained by the system are between 40% and 60%, as shown in Fig. 4.6 (Bottom) and Fig. 4.7 (Bottom). It is interesting to note that system performance was much more stable across participants than human performance. In addition, in terms of phoneme units, the global mean of the automatic system was 51.25%, very close to the global mean of 52.20% obtained by humans.

There are several factors that help understanding why the system achieves significantly higher rates in terms of phonemes than in terms of words: 1) Phoneme accuracy is computed at frame level because that is the output rate of the system. Thus, the temporal resolution used for phonemes is much higher than that of words and correctly recognizing a word implies the correct match of a rather long sequence of contiguous phonemes. Any phoneme mismatch, even if in a single frame, results in the whole word being wrong. 2) The automatic system finds it easier to recognize concrete phonemes (e.g: vowels) with high appearance rates in terms of frames (vowels are usually longer than consonants). This implies that a high phoneme recognition rate does not necessarily mean that the message is correctly decoded. To analyze this, system performance is displayed in Fig. 4.8. Specifically, in Fig. 4.8 (Top) we can observe the number of phonemes that were wrongly detected, distinguishing false negatives (in red color) and false

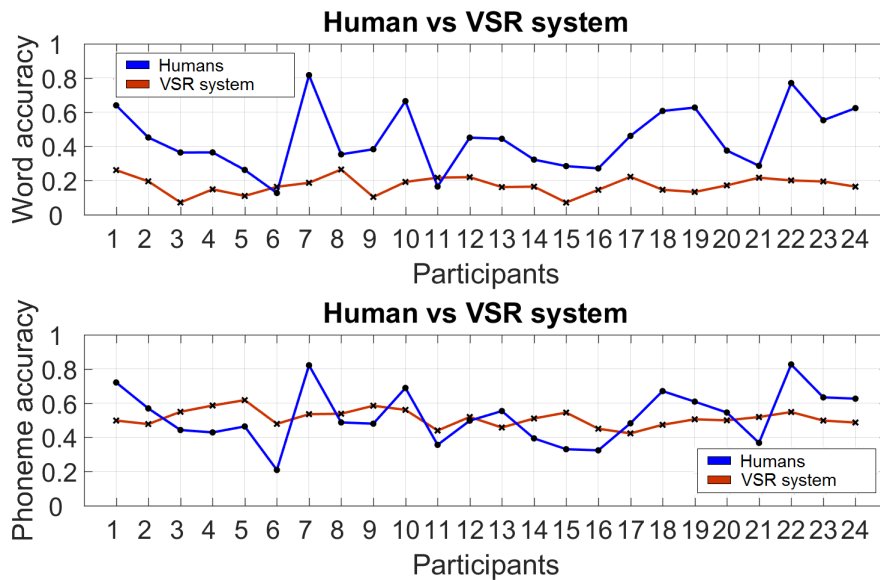


Figure 4.7: Top: human observers performance (Repetition 1) and automatic system performance for each participant in terms of word recognition average; Bottom: human observers performance (Repetition 1) and automatic system performance for each participant in terms of phoneme recognition average.

positives (in green), while Fig. 4.8 (Bottom) shows the corresponding values of precision and recall. Most of the consonants have very high precision, but many samples are not detected, deriving in a low recall. In contrast, vowels have an intermediate precision and recall because they are assigned more times than their actual occurrence. Close inspection of our data suggests that this effect is partially explained by the difficulty in correctly identifying the temporal limits of phonemes.

4.4 Discussion and Conclusions

In this work we explore visual speech reading with the aim to estimate the recognition rates achievable by human observers and by an automatic system under optimal and directly comparable conditions. To this end, we recorded the VLRF database, appropriately designed to be visually informative of the spoken message. For this purpose we recruited 9 hearing-impaired and 15 normal-hearing subjects. Overall, the word recognition rate achieved by the 24 human observers ranged from 44% (when the sentence was pronounced only once) to 73% (when allowing up to 3 repetitions). These results are compatible to those from Duchnowski et al. [50], who stated that even under the most favorable

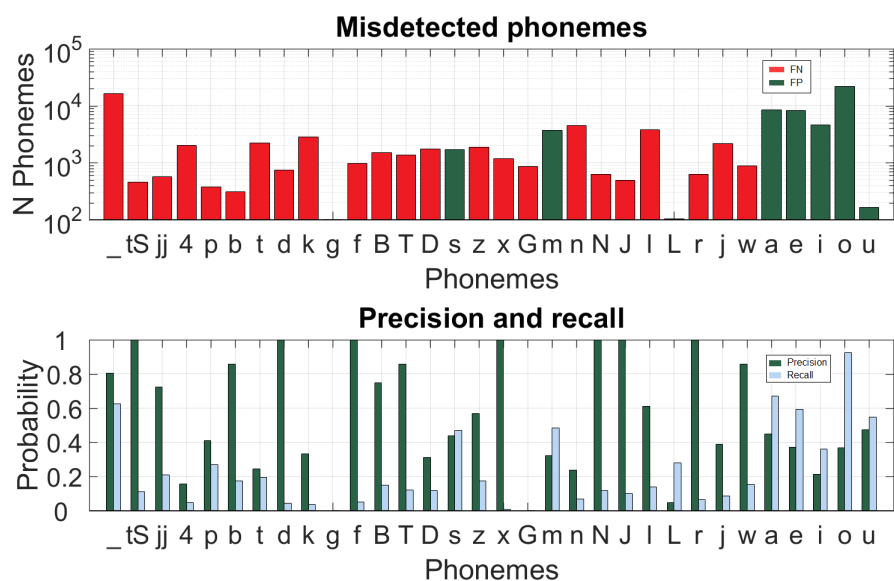


Figure 4.8: Top: Number of wrong detected phonemes. The red columns represent the false negatives phonemes and the green ones the false positives.; Bottom: Precision and Recall of each phoneme.

conditions (including repetitions) ”speech-readers typically miss more than one third of the words spoken”.

We also tested the performance of participants grouped by their hearing condition to compare their lip-reading abilities and verify if these are superior for hearing-impaired subjects, as suggested in some studies. Concretely, we found that hearing-impaired participants outperformed normal-hearing participants on the lip-reading task, but without statistical significance. The performance difference, which averaged 20%, was not sufficient to conclude significance with the current number of subjects. Hence, future work will address the extension of the VLRF database so that it includes sufficient subjects to reach a clearer conclusion.

The participation of hearing-impaired people was very important given their daily experience in lip-reading. During the recording sessions they explained that lip-reading in our database was a challenge because they did not know the context of the sentence beforehand. For them, it is easier to lip-read when they know the context of the conversation. The conversation topic constrains the vocabulary that can appear in the talk. Furthermore, we mentioned before that lip-reading is related to intelligence and language knowledge. During the recording sessions we noticed that sentences directly related to daily life were easier to understand than sentences with words not used in colloquial language.

Another important aspect to consider is how easy or difficult is to lip-read different speakers. As explained in Section 4.2, participants were instructed to use their own criterion to facilitate lip-reading. It is difficult to objectively judge the effectiveness of the techniques that were used, but we observed some interesting tendencies during the recordings. Firstly, facial expressions help decoding the spoken message adding context to the sentence (e.g: sad expression if you are speaking about something unfortunate); hearing-impaired participants used this technique more often than normal-hearing subjects. Secondly, it is more useful to separate clearly between words than to exaggerate pronunciation. That is because the human system is searching for words that fit the lip movements. We noticed that when pronunciation was exaggerated the separation between words was not clear or even lost considerably increasing the difficulty of lip-reading.

The above is important when interpreting the results of human observers, as they are conditioned both by the lip-reading abilities of the lip-reader and by the pronunciation abilities of the speaker. Recall that, in our experiments, each participant only lip-read his/her corresponding partner. It would be interesting to separate these factors, which could be done by randomizing the combinations of speakers and lip-readers on a per-sentence basis. In particular, the most interesting aspect would be to estimate the level of difficulty to lip-read each of the speakers, which could be done by having several subjects lip-reading the same speaker. There would be several advantages in doing so: 1) it would allow a more direct comparison to the performance of the system, as speaker performance will not be conditioned to a single human reader; 2) speakers that are too difficult could be excluded from the analysis, at least when seeking for the theoretical limit of lip-reading in optimal conditions; 3) it would help understand which are the best speaking techniques to use to facilitate lip-reading understanding.

As just explained, in our experiments, human observers reached word accuracy of 44% in the first attempt while our visual-only automatic system achieved 20% of word recognition rate. However, if we repeat the comparison in terms of phonemes, the automatic system achieves recognition rates quite similar to human observers, just above 50%. These results are comparable with those reported by Lan et al. [113] who tested in the RM corpus, using 12 speakers and 6 expert lip-readers. Concretely, their human lip-readers reached 52.63% viseme accuracy (in our case 52.20% phoneme accuracy) and their system obtained 46% viseme accuracy (our system 51.25% phoneme accuracy). Therefore, in terms of viseme/phoneme accuracy, both Lan's and our system reach near-human performance. But this does not happen in terms of word accuracy: Lan et al. reported human word accuracy of 21% (ours 44%) and system word accuracy of 14% (ours 20%).

When trying to explain the above, we found that the low word recognition rates were related to: 1) the fact that it is quite easy to make mistakes at frame

level and a mistake in a single frame results in the whole word being wrong; 2) the imbalance in the occurrence frequencies of phonemes. The latter is especially important because it highlights that the system, while achieving similar phoneme rates to those from humans, does not actually perform equally well. In other words, the phoneme sequences returned by humans always make some sense, which is not generally true for the system as it does not include higher-level constraints (e.g. at the word- or phrase-level). Hence, future directions should focus on introducing constraints related to bigger speech structures such as connected phonemes, syllables or words.

Chapter 5

END-TO-END LIP-READING WITHOUT LARGE-SCALE DATA

There is an increasing interest in interpreting speech using only visual information, which has led to growing research efforts on the development of Automatic Lip-Reading (ALR) systems that can work on realistic application settings, i.e. *continuous lip-reading*. Thus, in the last years there has been a progressive shift from simpler and constrained recognition tasks such as digits or letters recognition to more complex and natural scenarios such as words, sentences, or continuous speech recognition [62]. An important factor for this development has been the emergence of Deep Neural Networks (DNN), which have significantly pushed forward the achievable performance in ALR, though at the expense of requiring massive amounts of training data.

End-to-end DNNs consist of several hidden layers with millions of parameters between the input and the output that are capable of addressing complicated classification tasks working directly on the raw input data. Because of their large number of parameters, such models are data-hungry, and an extensive number of representative samples are required for their training, i.e. the *larger* the training set, the *better* the performance of the algorithm. Thus, in order to properly train end-to-end DNNs, we must find a balance between the amount of available training data and the number of parameters of the model.

Unfortunately, in lip-reading, data been so far an important limitation, especially for languages different from English. Most audio-visual databases suitable for ALR are not sufficiently large or do not cover enough vocabulary to train end-to-end architectures that generalize well. To illustrate this, Table

Adapted from: Fernandez-Lopez, A., & Sukno, F. M. (2020). End-to-end Lip-Reading without Large-Scale Data. *International Journal of Computer Vision*. (Under Review)

Table 5.1: Some of the largest audio-visual databases for continuous speech recognition for various languages.

Dataset	Language	Utterances	Hours
LRS2-BBC	English	144k	225
LRS3-TED	English	152k	438
LSVSR [†]	English	2.9M	3886
UWB-07-ICAV	Czech	10k	25
NDUTAVSC	German	6k	11
BL	French	4k	6
HAVRUS	Russian	4k	6
VLRF	Spanish	600	3
Wild LRRo [‡]	Romanian	1k	21
Lab LRRo [‡]	Romanian	8k	5

[†] Non-publicly available

[‡] Dataset in words

5.1 shows examples of audiovisual datasets for continuous ALR in different languages. From them, we observe that a few recent corpora recorded in English contain more than 400 hours of video recordings with annotated transcriptions and thousands of training sentences. In contrast, such large-scale datasets cannot be found in any other language. For instance, for two widely spoken languages such as Spanish and French, the available datasets are considerably smaller, with 600 and 4000 utterances for training continuous ALR systems, respectively. Therefore, these datasets are more than 100 times smaller than the largest English-spoken ones, which makes it very difficult to train competitive end-to-end DNNs that are comparable to the state-of-the-art systems presented in English [11]; [41]; [202]; [134]; [4]. Moreover, the acquisition of new databases is challenging and time-consuming, especially due to the need for appropriate labeling (e.g. text or phonemes aligned with the video stream), which is tedious, time-consuming and error-prone. Therefore, a system that could be trained on small-scale datasets and still achieve competitive performance would be very beneficial for ALR.

Related work: The design of end-to-end architectures for small-scale databases has mainly followed two strategies: a) to use pre-trained models to avoid having to train DNNs from scratch; b) to deal with low resource data designing alternative architectures or Data Augmentation (DA) techniques. Among systems that use pre-trained models, some authors have explored the use of pre-trained networks designed for other computer vision applications, e.g. AlexNet, VGG, GoogLeNet or ResNet [196], but their results are significantly below the models

specifically trained for lip-reading [43]. In contrast, other researchers deal with small-scale datasets (e.g. OuluVS2, CUAVE) without using external data to train their models [65]; [172]; [71]; [117]. For example, [71] proposed an end-to-end DNN model where DA was crucial to circumvent the issue of insufficient training data. On the other hand, [177] proposed an encoding network combined with unidirectional and bidirectional recurrent network where the encoding layers were pre-trained in a greedy layer-wise manner using Restricted Boltzmann Machines (RBM) to avoid adding external data. However, those ALR systems target word or sentence recognition tasks in constrained scenarios using datasets such as OuluVS2 or GRID, where the output of the system is restricted to a pre-defined number of possible classes. While this is useful to analyze the effectiveness of algorithms at early design stages, the resulting models tend to be of limited scope and difficult to extrapolate to more complex tasks such as natural speech. In contrast, we are looking for lip-reading systems that handle more complex and realistic settings targeting continuous lip-reading which means that the system must be able to decode any word of the dictionary and process sentences that contain an arbitrary number of words with unknown time-boundaries.

Contribution: in this work, differently from previous approaches, we show that it is possible to train competitive end-to-end ALR systems with challenging small-scale datasets as long as the appropriate restrictions are made to the learning process, especially in terms of the visual front-end objective. To this end, we revisit the convenience of targeting the standard *phonemes* (defined as the minimum distinguishable acoustic units that are able to change the meaning of a word) or the controversial *visemes* (the visual domain equivalent) [222]; [66], and hypothesize that the visual front-end should instead be trained in a self-supervised setting, allowing it to target its own *visual units*, which we define as *a collection of visually similar images constrained by linguistics*. We translate this definition into a mathematical formulation based on simple constraints and show that these visual units can be used to add an intermediate classification task between the visual and temporal modules that facilitates meaningful learning of visual features and, as a consequence, reduces the amount of data required to train a standard end-to-end architecture consisting of a CNN-based visual module followed by an attention-based sequence-to-sequence module that predicts continuous speech in terms of characters.

Additionally, we also present a data augmentation strategy that allows synthesizing novel realistic video sequences by appropriately combining characters-like sub-sequences from existing videos, and find that this allows enriching the temporal context learned by the sequence-to-sequence module. We test the proposed system on the VLRF dataset [61], a small-scale database that is however one of the largest in Spanish, and achieve 44.77% CER and 72.90% WER, which are competitive with the state-of-the-art and arguably the best to date

for this volume of training material.

5.1 Training with limited data

If we attempt to train end-to-end ALR systems for any language but English, we soon realize that we are short of data, since there is considerable imbalance between the amount of available training data and the number of parameters of the model (see Table 5.1). For example, the well-known AlexNet for object classification contains 62 millions of parameters, which were trained from 14 million images ($\sim 22\%$ of the number of parameters). In contrast, the number of parameters of current ALR systems (based on combinations of CNN and LSTM networks) is in the order of millions (~ 12 millions) while the amount of available training samples using most audio-visual lip-reading datasets (which can be considered small-scale), can be substantially reduced. For example, when using the VLR dataset, which consists of 600 sequences averaging 5 seconds each or the BL dataset, which consists of 4,000 sentences averaging 2 seconds each, we do not reach even a 2% of data with respect to the number of parameters for training, which makes it very difficult and time-consuming to train the network at once, even considering data augmentation techniques.

Consequently, instead of training the whole system for the single task of speech recognition at the character-level, we propose to add an intermediate task between the visual and the temporal modules, which is related to mouth position classification. It is quite intuitive to add an intermediate classification task in this way because each module has a specific goal that could be reached jointly or independently. The goal of the visual module is to parametrize the visual information observable at a given time instant or window (i.e. analog to phonemes in speech). On the other hand, the aim of the temporal module is to map the visual features into speech units while incorporating temporal constraints to ensure that the decoded message is coherent. Therefore, we propose to divide the whole set of sentences into small speech units to adequately constrain the training of the visual module. In this way, we will be able to control the network learning to ensure that the extracted features are representative enough to appropriately encode the mouth appearance in a way that is helpful for the temporal module in order to predict the character.

To do so, we would ideally need a labeled dataset that provides very accurate speech labels, i.e. phonemes or visemes. Unfortunately, most of the lip-reading datasets provide only the text that corresponds to each phrase but does not provide phoneme or viseme labels per frame. Furthermore, while there exist semi-automatic programs such as Praat [24] or Montreal Forced Aligner [141] to align the text and the audio stream, they often require considerable manual intervention

to refine the boundaries of each phoneme, resulting in a challenging and time-consuming process which does not scale well.

As a solution, based on the observation that the visual module only needs to distinguish among visually separable classes, we propose to rely on weak labels that can be easily obtained in a fully automatic manner but are still informative about the mouth appearance. Hence, we hypothesize that, if the CNN is able to differentiate among visually separable classes, the features generated at the last step of the visual module (those of the fully-connected layers) will properly encode the mouth appearance and will be helpful for the temporal module to decode visual speech and predict the correct character.

5.1.1 Visual Units

5.1.1.1 Motivation and definition

At first glance, one may think that it would be ideal if the training set for the visual module contained accurate speech labels for each input frame. However, apart from being impractical due to the amount of human supervision required for the task, the selection of the labels that should be used is also not trivial. It is widely accepted that phoneme labels can be ambiguous in a visual speech setting because i) phonemes like */b/* or */p/* are visually indistinguishable; ii) phonemes like */k/* can be produced with quite different mouth positions depending on the preceding or following sounds. On the other hand, the definition of the visual equivalent to phonemes (visemes) is still an open problem, lacking of a standard and with considerable controversy [18]; [87]; [183]; [152].

Thus, the aim of a large scale supervised setting with highly accurate labels is unlikely. In contrast, the automatic generation of weak labels to enforce the training of the visual module inherently eliminates the need for human labeling and facilitates obtaining larger training sets. Nevertheless, a crucial aspect of such an approach is whether the automatic labels would be sufficiently informative to correctly train the visual module. In this sense, it is important to realize that:

- We cannot aim at automatic labels that perfectly decode speech, since that would mean trying to completely solve the visual speech recognition problem by means of a simple CNN-based visual module, disregarding temporal context.
- Even if the visual module cannot perfectly decode speech, it is reasonable to aim at extracting features that encode the appearance of the region of interest (i.e. the facial area around the mouth and the lips).

Therefore, we will not target a visual module that can decode speech, but a visual module that produces features that are informative about the appearance of

the mouth and the lips. To train such a visual module we do not need phoneme labels: we need labels that indicate when the mouth and lips in one frame are similar or different from the mouth and lips in another frame, together with constraints that allow establishing some relation between speech and the labels assigned to those frames.

Thus, we define visual units as *a collection of visually similar images constrained by linguistics*, based on the following constraints:

1. The similarity between images labeled with the same visual unit must be computed excluding inter-subject differences.
2. Labeling into visual units should induce the segmentation of a video sequence in groups of consecutive frames that share the same label.
3. The above segments should be related to the phrase or sentence uttered in the video sequence, much like a mapping from visual units to speech units (i.e. characters, phonemes, etc).
4. Visual units should not be subject-specific, but common to a large number of subjects (not necessarily to all subjects, since not all speakers necessarily use exactly the same pronunciation units).

In the next paragraphs, we provide a mathematical formulation to define visual units based on the above constraints. Later, in Section 5.2, we describe the algorithmic implementation used in this paper to derive the visual units that will be used in our experiments.

5.1.1.2 Formulation

We propose to automatically generate weak frame labels to constrain the training of the visual module by minimizing the energy functions presented below in (5.1) and (5.6). We define a function $b: \mathbb{Z} \rightarrow \mathbb{Z}$ that maps a frame m from subject s into a visual unit $v \in \mathcal{V} = \{1, \dots, V\}$, where \mathcal{V} is the set of visual units with length V . For the rest of the paper, $b(m)$ will be denoted as b_m for simplicity.

In the following subsections, we explain step by step the proposed energy functions where (5.1) generates a set of visual units specific to each subject $\mathcal{V}^s = \{1, \dots, V^s\}$ and (5.6) finds a common set of visual patterns among users that generalizes the speaker-dependent sets of visual units $\mathcal{V}^s \forall s$ into a global set of visual units \mathcal{V} .

Speaker-specific visual units Minimization of equation (5.1) produces the set of visual units \mathcal{V}^s for each speaker s . Intuitively, the first term in (5.1) derives directly from the definition of visual units, which shall be groups of frames with

$$\begin{aligned}
\forall s \in \mathcal{S} : V^s = \arg \min_k & \sum_m \sum_{n \neq m} \|I_m - I_n\|_2 \delta[b_m - b_n] \\
& + \lambda_1 \sum_m \sum_{\substack{n \neq m \\ |m-n| \leq W}} \phi_a(\|I_m - I_n\|) \phi_t(\|m - n\|) (1 - \delta[b_m - b_n]) \\
& + \lambda_2 \sum_u \sum_k (\|P_u - T_{u,k}\|_2 + \lambda_3 \|P_u^\dagger - Q_{u,k}\|_2)
\end{aligned} \tag{5.1}$$

a similar appearance. However, frame similarity is a multifactorial attribute, i.e. two frames m and n can be similar because they share the the same mouth position ($I_m \sim I_n, b_m = b_n = v$) or the same subject identity ($I_m \sim I_n, m, n \in s$). To factor-out the latter, at this stage we enforce subject-specific visual units v , which shall be composed by groups of frames with similar lip-positions. To achieve this, we penalize that two frames m and n of the same subject $s \in \mathcal{S} = \{1, \dots, S\}$ get assigned to the same visual unit v if there are large intensity differences between them, with $\delta(\cdot)$ being the Kronecker delta and S the number of subjects.

The second term in (5.1) controls temporal coherence. It is assumed that neighboring frames (with a maximum distance W) should correspond to the same visual unit v , unless they have a large appearance difference. Thus, we enforce that frames m and n that are temporally close and have similar appearances are assigned the same visual unit. We do so by penalizing with $1 - \delta[b_m - b_n]$ weighted by a spatio-temporal bilateral filter that depends on both the appearance difference ϕ_a and the temporal distance ϕ_t (i.e. ϕ_a and ϕ_t are Gaussian kernels). As we will show later, these temporal constraints induce temporal segments of consecutive frames that are labeled with the same visual unit until a new appearance transition occurs.

The third term in (5.1) controls the number of visual units per speaker V^s . To determine V^s we compare the segments induced by our visual units with respect to the speech transcript of every training sentence. Considering that we are looking for visual units that are informative about speech, we expect that the number of visual unit segments in a given sentence is similar to the number of speech units therein; and also that repetitions of the same speech unit result in repetitions of the same visual units. In other words, we would ideally expect that a video sequence uttering the word "casa" would be labeled into 4 visual unit segments ($c-a-s-a$), two of which would share the same label. In practice, the relation between mouth movements and speech is more complex than this idealized example, but we can still use the same intuition to establish V^s . Specifically, for each training sentence, we look for a balance between visual units variability, i.e. the number of different

visual units observed in the sentence; and visual units continuity, i.e. the number of temporal segments that are induced in the sentence.

To find the above balance we compare the segments induced by different numbers of visual units, indexed by k in (5.1). Let us define P_u as the total number of phonemes P that are spoken in the utterance u , and from there $P_u^\dagger \leq P_u$ as the number of unique phonemes. Concretely, for each set of k visual units, we minimize the distance between the number of induced time-segments and the number of real phonemes ($T_{u,k} \sim P_u$); and the distance between the number of different visual unit labels and the number of unique phonemes ($Q_{u,k} \sim P_u^\dagger$). Finally, the optimal number of visual units per subject V^s will be the k^{th} set that jointly minimizes both distances.

Fig. 5.1 illustrates the temporal segments induced for the utterance "Miraba el reloj" by different sets of visual units. This sentence consists of $P_u = 13$ time-segments and $P_u^\dagger = 9$ different sounds. To illustrate this example, imagine that this utterance u contains 26 frames. Then, we observe that with a set of $k = 15$ visual units we obtained a mapping that produces $T_{u,15} = 11$ time-segments and $Q_{u,15} = 8$ different visual units from the set of 15. In contrast, we observe that with a set of $k = 11$ visual units we obtain a mapping that produces $T_{u,11} = 8$ time-segments and $Q_{u,11} = 5$ visual units. In this particular example, the set $k = 15$ is selected because it generates a sequence that is more consistent with the text transcript, i.e. $\|T_{u,15} - P_u\| < \|T_{u,11} - P_u\|$ ($\|11 - 13\| < \|8 - 13\|$) and $\|Q_{u,15} - P_u^\dagger\| < \|Q_{u,11} - P_u^\dagger\|$ ($\|8 - 9\| < \|5 - 9\|$).

Generalization to a common set Once that we have all sets of speaker-specific visual unit, we aim to achieve a mapping from any frame m into a visual unit v that is independent of the subject (i.e. $\forall m, b_m \rightarrow v \in \mathcal{V} = \{1, \dots, V\}$). To do so, we wish to merge all the speaker-specific sets of visual units \mathcal{V}^s into a global set of speaker-independent visual units, \mathcal{V} . However, the mapping between all speaker-specific sets is far from trivial, because:

1. Even though most of the pronunciation patterns are common among users, the number of visual units V^s can change depending on the speaker because every person pronounces in a unique way and there are people that have a larger visual speech variability than others, i.e. they use more lip movements when they speak than others [119, 243]. Thus, a one-to-one mapping between all pairs of subjects is unlikely.
2. As illustrated in Fig. 5.1, the segments induced by visual units cannot be assumed to directly correspond to a character or to a phoneme. Recall that this is a key aspect of visual units, which are defined primarily in terms of visual appearance rather than in terms of speech, although they are expected to be informative also about the latter.

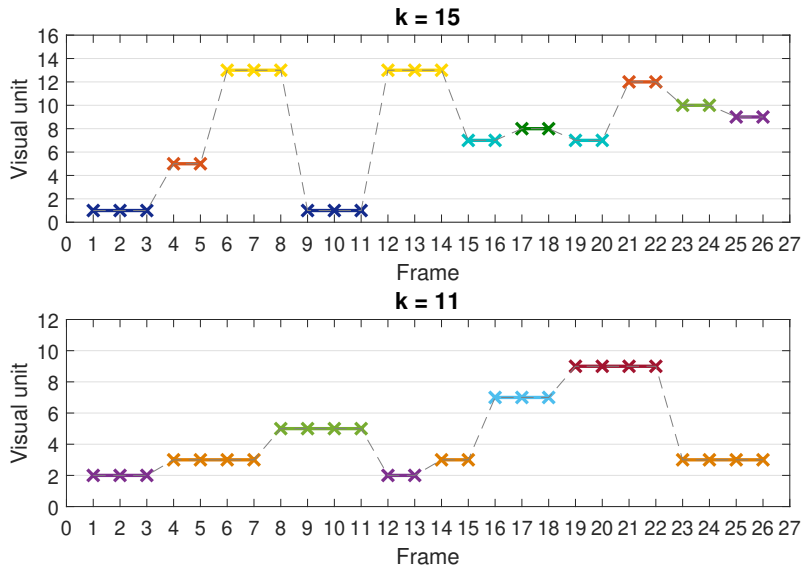


Figure 5.1: Mapping into visual units of the phrase "Miraba el reloj" for sets with 15 and 11 visual units.

3. In the general case, the training sentences for each speaker can vary and might even be unique to some speakers, thus discouraging the short-cut of finding repeated sentences to relate the speaker-dependent visual units.

Therefore, we need a mechanism to establish correspondences between the visual units from different speakers, each of them uttering a possibly different set of sentences. For this purpose, we propose to rely on an estimate of the phonetic distribution associated with each visual unit. For the sake of argument, assume that for each frame in the input sequences there is a phonetic label associated with it (e.g. a ground-truth segmentation into phonemes)¹. Because each frame is also labeled in terms of speaker-specific visual units, we can estimate a probability density function (pdf) that represents the phonetic distribution associated with each visual unit (for each speaker). In this way, we can address the search for correspondences based on the assumption that, if two visual units from different speakers correspond to the same visual pattern, then they should have similar phonetic distributions.

Let us define the probability distribution of each visual unit v^s as a p -bin histogram that measures the frequency of occurrence of the phonemes set $\mathcal{P} =$

¹In our experimental evaluation we will show that this requirement can be relaxed to work directly from very rough estimates of the phonetic segmentation with very little impact on the final accuracy.

$\{1 \dots p\}$. Then, we denote the estimated probability distribution as

$$\hat{\mathbf{q}}(v^s) = \{\hat{q}_i(v^s)\}_{i=1 \dots p} \quad \sum_{i=1}^p \hat{q}_i(v^s) = 1 \quad (5.2)$$

The phonetic information of each visual unit can be used to directly compare two different visual units (v^s and $v^{s'}$, $s \neq s'$) by computing the similarity between their pdfs. Thus, the similarity function would define the distance between two visual units from different speakers. In this case, we define the distance between two discrete phonetic distributions of visual units v^s and $v^{s'}$ from different speakers s and s' as

$$d(v^s, v^{s'}) = \sqrt{1 - \rho(v^s, v^{s'})} \quad (5.3)$$

where we chose

$$\rho(v^s, v^{s'}) = \sqrt{\hat{\mathbf{q}}(v^s) \cdot \hat{\mathbf{q}}(v^{s'})} \quad (5.4)$$

the sample estimate of the Bhattacharyya coefficient between $\hat{\mathbf{q}}(v^s)$ and $\hat{\mathbf{q}}(v^{s'})$ [100], where $\hat{\mathbf{q}}(v^s)$ is the estimate of the phonetic distribution of visual unit v from subject s ($v^s \in \mathcal{V}^s$) and $\hat{\mathbf{q}}(v^{s'})$ is the estimate of the phonetic distribution of visual unit v from any other subject s' , $s' \neq s$ and $v^{s'} \in \mathcal{V}^{s'}$.

The function $\rho(v^s, v^{s'})$ plays the role of a likelihood and its local maximum indicates two visual units are candidates to be in correspondence. Then, given a visual unit v^s from subject s and the whole set of visual units $\mathcal{V}^{s'}$ from a different subject s' , we define $F(v^s, \mathcal{V}^{s'})$ in (5.5) as the function that minimizes the phonetic distance and returns the most similar visual unit $v^{s'}$, i.e. $\hat{\mathbf{q}}(v^s) \sim \hat{\mathbf{q}}(v^{s'})$. To ensure reliable correspondences, we consider that two visual units v^s and $v^{s'}$ from different subjects s and s' are in correspondence if and only if the relationship obtained from (5.5) is invariant to a swap between s and s' , that is, if we obtain v^s when minimizing $F(v^{s'}, \mathcal{V}^s)$ and also $v^{s'}$ when minimizing $F(v^s, \mathcal{V}^{s'})$.

$$F(v^s, \mathcal{V}^{s'}) = \arg \min_{v^{s'}} \left(\sum_{v^{s'} \in \mathcal{V}^{s'}} \|d(v^s, v^{s'})\| \right) \quad (5.5)$$

Then, we define $\mathbf{Z}(s, v^s, s')$ in (5.6) as the function that minimizes the phonetic distance between visual unit v^s from subject s and all visual units from subject s' and finds (if it exists) the visual unit $v^{s'}$ that is in correspondence with v^s .

$$\mathbf{Z}(s, v^s, s') = \begin{cases} v^{s'} & \text{if } v^s = F(F(v^s, \mathcal{V}^{s'}), \mathcal{V}^s) \\ 0 & \text{otherwise} \end{cases} \quad (5.6)$$

Calculation of $Z(s, v^s, s') \forall s$ and $\forall s'$ helps to determine the set of visual units that are in correspondence with each v^s . Tracking this correspondence between visual units across subjects, we can merge those visual units that are common to groups of subjects until all speaker-dependent visual units v^s have been assigned into a speaker-independent set \mathcal{V} . Specifically, we do this following Algorithm 1, in which we define $J(s, v^s)$ in (5.7) as the number of visual units $v^{s'}$ that are in correspondence with v^s . Thus, $J(s, v^s)$ allows us to iteratively merge those visual units that are common to a larger number of subjects (in descending order), i.e. the most common visual units are firstly assigned to \mathcal{V} repeatedly until all visual units v^s have been assigned. We also define $A(s, v^s)$ as a matrix that controls if a visual unit v^s has been already assigned to v .

$$J(s, v^s) = \sum_{\forall s'} \mathbb{1}[Z(s, v^s, s') > 0] \quad (5.7)$$

In this way, we finally obtain our global set \mathcal{V} and consequently our direct mapping from any frame m into a speaker-independent visual unit v , i.e. $\forall m, b_m \rightarrow v, v \in \mathcal{V} = \{1, \dots, V\}$.

Algorithm 1 Generalization to a common set \mathcal{V}

```

Input  $Z, J$ 
Initialize  $\mathcal{V} = \emptyset$  and  $A = \emptyset$ 
for  $t = S - 1$  to  $1$  do
  for  $s = 1$  to  $S$  do
    for  $v^s = 1$  to  $V^s$  do
      if  $J(s, v^s) \geq t$  then
        if  $A(s, v^s) == \emptyset$  then
          Add new visual unit  $v$  in  $\mathcal{V}$ 
          Label all frames from  $v^s$  as  $v$ 
           $A(s, v^s) = v$ 
        for  $s' = 1$  to  $S, s' \neq s$  do
           $v^{s'} = Z(s, v^s, s')$ 
          if  $Z(s', v^{s'}, s) == v^s$  then
            if  $A(s', v^{s'}) == \emptyset$  then
              Label all frames from  $v^{s'}$  as  $v$ 
               $A(s', v^{s'}) = v$ 

```

Output A

5.2 Frame mapping into visual units

In this section, we describe the algorithmic implementation that we will use in this paper to derive visual units (defined in Section 5.1.1).

Let \mathcal{X} be the data to train our ALR system, where each sample of \mathcal{X} comprises a variable-length video sequence $\{I_1, I_2, \dots, I_M\}$ and its corresponding grapheme transcription $\{y_1, y_2, \dots, y_L\}$; typically $L < M$. All those samples belong to a set of utterances $u \in \mathcal{U} = \{1, \dots, U\}$ of different speakers $s \in \mathcal{S} = \{1, \dots, S\}$. Our goal in this section will be to derive a new set $\tilde{\mathcal{X}}$ to pre-train our visual front-end. Specifically, once the visual units are defined, each sample of $\tilde{\mathcal{X}}$ will correspond to the original set of video frames in \mathcal{X} , namely $\{I_1, I_2, \dots, I_M\}$ but now with an associated set of visual unit annotations $\{v_1, v_2, \dots, v_M\}$, where each $v \in \mathcal{V} = \{1, \dots, V\}$.

To obtain $\tilde{\mathcal{X}}$ we follow the definition of visual units introduced in Section 5.1.1.1, which we divide in 3 steps:

1. To enforce that neighboring frames with similar mouth appearance get assigned to the same visual unit, we introduce a deep autoencoder that can be trained to jointly minimize the first and second terms of equation (5.1). This yields a subject-specific latent space in which samples are arranged according to the spatio-temporal constraints just mentioned (Section 5.2.1).
2. To convert the above latent representation into visual units, we address the minimization of the third term in equation (5.1), which aims to relate visual features to speech. Because the first two terms are already fixed, this can be done by exhaustive search of the optimal number of visual units per subject (Section 5.2.2).
3. Finally, subject-specific visual units are merged into a common set, which is now subject-independent (Section 5.2.3).

5.2.1 Deep latent features

We map the input images from each subject into a low-dimensional representation that enforces similarity between samples that are close in time and show similar lip position, while also enforces dissimilarity between samples that show a different lip position. Inspired by [8], we propose to do this by means of a deep clustering method that consists of a Deep Convolutional Auto-Encoder (DCAE) trained with triplet loss and followed by k -means clustering.

The use of the Triplet-Loss is critical to appropriately enforce the spatio-temporal constraints that we target. Given an *anchor* sample, the encoder projects it into a low-dimensional space in which it minimizes the distance between the

representation of the anchor sample and a visually similar sample (*positive*) and maximizes the distance between the input sample and a visually different sample (*negative*). Thus, positive samples will be close in time and appearance to the anchor sample, while negative samples will show a different mouth appearance.

We use the Mean Absolute Difference (MAD) to quantify how similar is the appearance between two frames. Specifically, given an anchor frame I_m , we compute its visual dissimilarity with respect to all other frames in a given utterance u with:

$$D(m, n) = \frac{1}{N} \sum_{x,y} |I_n(x, y) - I_m(x, y)| \quad (5.8)$$

$$\forall n \neq m, \quad \{I_m, I_n\} \in u$$

and define what frames can be used as positive and negative samples based on statistics from $D(m, n)$. Firstly, positive samples are defined as those frames in the utterance that are significantly more similar to the anchor I_m than the rest. This is done by defining an outlier threshold based on the quartiles of $D(m, n)$, denoted by $\{q_{25}, q_{50}, q_{75}\}$, which are computed in all frames except those in a small neighbourhood W of the anchor I_m . The latter is necessary due to the high similarity of neighbouring frames, which could unreasonably reduce the estimated values of the quartiles. Once the quartiles are estimated, all frames with similarity below the standard outlier threshold of $q_{50} - 1.5 \times (q_{75} - q_{25})$ are considered suitable *positive samples*.

In contrast, any frame not selected as an outlier could, in principle, be considered as a potential negative sample. However, it is important to find a balance between easy and difficult cases [90], so that the encoder can be efficiently trained. In our experiments, we found the median to be an appropriate threshold; hence, all frames I_m with $D(m, n) \geq q_{50}$ are considered suitable negative samples. This allows both hard and easy negative samples, being *hard* those samples that are close to the median and *easy* those that are far above the median.

Fig. 5.2 illustrates this with an example taken from the VLR database. We show the resulting MAD with respect to reference frames I_{60} and I_{80} . In this figure, we observe that many frames will be considered negative samples while only a few, which are very close to I_m , are similar enough to be considered positives.

For a given anchor I_a , with positive and negative samples I_p and I_n , the triplet loss is defined as follows:

$$L_{Triplet}(I_a, I_p, I_n) = \max \{ \|f(I_a) - f(I_p)\|_2 - \|f(I_a) - f(I_n)\|_2 + \alpha, 0 \} \quad (5.9)$$

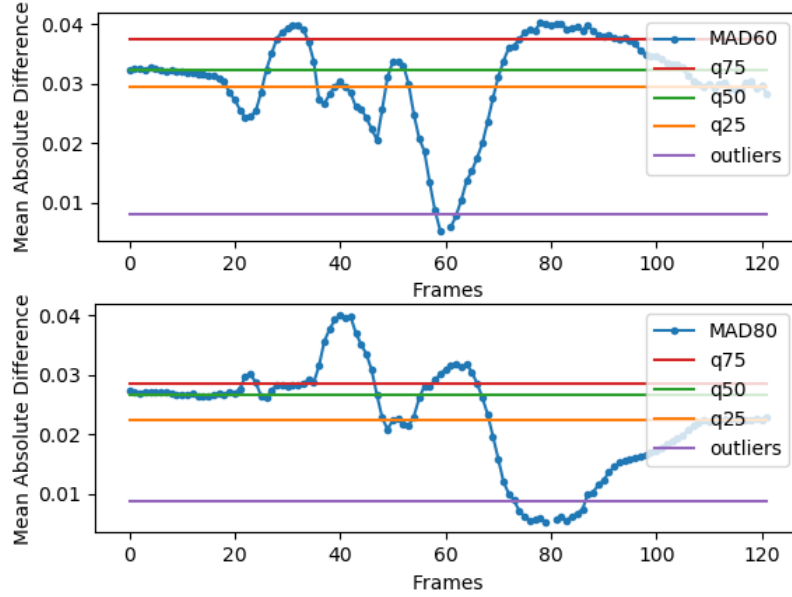


Figure 5.2: Example of MAD for frames $m = 60$ and $m = 80$ for the phrase "Miraba el reloj".

The way in which we select our positive and negative samples inherently helps in the minimization of the first and second terms of (5.1). Specifically, when minimizing (5.9) we are enforcing the latent representation of two visually similar frames (I_a, I_p) to be close to each other, so that we can easily *cluster* them, while at the same time, we are enforcing the latent representation of two visually different samples (I_a, I_n) to be far from each other. Thus, using *DCAE* network we are able to project any input image into an adequate low-dimensional feature representation by minimizing both the reconstruction loss L_{MSE} and the triplet loss $L_{Triplet}$, as shown in (5.10).

$$L_{DCAE}(I_a, I_p, I_n) = \frac{1}{2}L_{Triplet}(I_a, I_p, I_n) + \frac{1}{6}L_{MSE}(I_a, \hat{I}_a) + \frac{1}{6}L_{MSE}(I_p, \hat{I}_p) + \frac{1}{6}L_{MSE}(I_n, \hat{I}_n) \quad (5.10)$$

5.2.2 Optimal number of visual units per subject

Once the deep embedding for each subject has been learned, we could map each frame m into a speaker-specific visual unit v^s using any clustering algorithm. However, we do not know how many distinct visual units are produced by each

speaker, i.e. the number of visual units V^s is unknown. Specifically, we assume that the optimal number of visual units V^s can be different for each speaker and determine it by exhaustive search. Since this search is comparatively cheap, we evaluate all values between an extremely reduced set of 5 visual units and an over-dimensioned set as large as the number of Spanish phonemes, $V^s = 30$.

Thus, for each speaker, we apply k -means on the learned latent space for $k \in [5, 30]$ and determine the optimal number of visual units as the value of k that best approximates the number of time-segments and distinct units per sentence, i.e. $T_{u,k} \sim P_u$ and $Q_{u,k} \sim P_u^\dagger$, thus minimizing the third term of (5.1). This process generates an independent set of visual units per subject \mathcal{V}^s , which needs to be generalized into a global set of visual units \mathcal{V} that is common to all subjects, as explained next.

5.2.3 Generalization of all speaker-specific sets into a common set

As a final step, we aim to achieve a mapping from any frame m into a visual unit v that is independent of the subject. Considering that there is no direct mapping between visual units across subjects, in Section 5.1.1.2 we proposed to estimate the speech distribution associated with each speaker-specific visual unit and assume that those visual units with similar phonetic distribution across subjects should be merged into the same visual unit v . However, the estimation of the phonetic distribution for each visual unit v^s would require the labeling of all input frames in terms of phonemes, which implies an undesirable manual pre-processing load. Fortunately, as will be shown in our experiments, the determination of visual units is not especially sensitive to such phoneme labeling. As a consequence, it is possible to derive visual units from approximate phonetic annotations. Specifically, recalling that each sample in our data X comprises a variable-length video sequence $\{I_1, I_2, \dots, I_M\}$ and its corresponding grapheme transcription $\{y_1, y_2, \dots, y_L\}$, we use a rough estimate of the phoneme label for each frame based on the mean duration of the phoneme.

We start by transcribing each grapheme sequence $\{y_1, y_2, \dots, y_L\}$ into a phonetic sequence $\{p_1, p_2, \dots, p_L\}$ using a grapheme-to-phoneme transcription tool or the grapheme-to-phoneme conversion rules own by the specific targeter language [181, 122]. The frequency of occurrence and mean duration of phonemes are widely available for several languages. Therefore, the phoneme transcription $\{p_1, p_2, \dots, p_L\}$, the mean duration of each phoneme, and the total duration of the sequences in frames M are enough to generate a rough estimate of phoneme labels per frame.

In Fig. 5.3, we show three examples of sequences in which we can observe

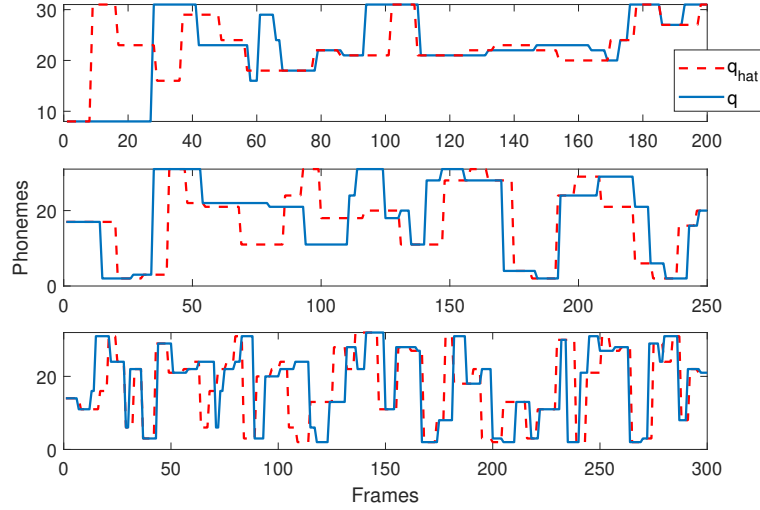


Figure 5.3: Example of approximate annotations per frame for 3 sequences.

the ground truth phonetic annotations per frame provided by the dataset (blue) and the approximate annotations per frame as described above (red).

Once a phonetic label per frame is available, it is straight forward to estimate the phonetic distribution of each speaker-specific visual unit as a p -bin histogram that measures the frequency of occurrence of the phonemes set $\mathcal{P} = \{1 \dots p\}$, following (5.2). We compute the phonetic distribution of each visual unit v^s and perform a one-by-one comparison of visual units across subjects, as it was shown in (5.4). Then, we merge those visual units from different speakers that are in correspondence as detailed in Algorithm 1. The merging process starts with those visual units that are common to the largest number of subjects, and proceeds in descending order as long as the visual units to merge comprise at least 10% of the speakers. Groups of visual units supported by less than 10% of the subjects are considered too specific and, therefore, are not included in the common set.

5.3 Spatio-temporal data augmentation

5.3.1 Motivation

The incorporation of DL techniques and the availability of large speech recognition datasets have pushed forward the achievable performance in speech recognition systems, which are capable of automatically transcribing spoken utterances with an accuracy above 95% when the signal is not corrupted by acoustic noise [39]. The same tendency can be observed in visual only speech

recognition systems, even though their recognition rates are still modest compared to those from audio systems. Recent trends suggest that the key for ALR relies on the proper modeling and interpretability of the context, in which DL has proven especially successful [62].

Speech context is defined as a continuous sequence of n different items from a given sample of speech. These items can be letters, phonemes, syllables, words, or even whole sentences. For example, consider a sequence of n phonemes, where its short context consists of its previous and following phonemes (tri-phonemes), while its long context consists of a long consecutive sequence of phonemes for every instant. We highlight the importance of both contexts because the short one, at the character level, helps in generating or decoding plausible words, while the long one, at the word and sentence levels, helps giving coherence to the message, filling the gaps or amending a miss-understood word by using its neighbors. DL techniques have shown to be very powerful to retain short and long term dependencies. Nevertheless, the success of these models relies on the availability of large amounts of training data, with sufficient variability at the different context levels. An ideal dataset should cover as many different words as possible, and combine them in many different phrases or sentences.

Unfortunately, very few datasets are as large and with so much variability as described above, and all of them are in English. For example, consider a moderate-size dataset such as VLRF, one of the largest ones in Spanish. It covers more than 10,000 words, but they are included in a comparatively small set of 600 sentences, which is quite limiting in terms of long-term context.

Thus, this is a challenging problem affecting most datasets available to date and all languages but English, which requires data augmentation techniques that enrich the context available for training. The first intuition could be to think about traditional data augmentation techniques in the spatial domain, (e.g. horizontal flips, rotations, shifts, zooming, ...), but they would not solve our problem because they would maintain the same linguistic content in terms of both semantics and syntax. Another alternative could be generative audio-visual speech synthesis, but current results seem to be still far from natural speech. Nevertheless, if the problem is constrained to a speaker-specific augmentation, we show below how we can re-use the available data to synthesize new sequences, allowing to increase the dataset context.

We based our synthesis scheme on the assumption that any plausible word can be generated using a combination of independent phoneme segments. In particular, the availability of a rich set of phonemes makes it possible to synthesize any feasible word or sentence by a simple combination of small speech units. For example, imagine that our set of phonemes comes from splitting the Spanish word *c-o-s-a* in its set of phonemes: *"/k/"*, *"/o/"*, *"/s/"*, *"/a/"*. Thus, following this statement, we could easily generate alternative words such as *a-s-c-o* or *c-a-s-*

o using a different combination of the same set of phonemes already available from *c-o-s-a*. In this way, we firstly propose to split our training set into small video sequences that contain the utterance of each phoneme; and secondly, to concatenate and interpolate those phonetic sequences to construct new phrases.

In the following subsections, we introduce the synthesis procedure that includes the data preprocessing and the assessment of synthesis plausibility.

5.3.2 Preprocessing the data

5.3.2.1 Collecting phonetic data for each speaker

Assuming an audio-visual dataset that provides phonetic labels, we propose to generate a subject-specific phonetic dataset that consists of consecutive frames uttering a single phoneme. Thus, for each speaker s and utterance u , we propose to split each sentence into the smallest speech units and annotate them as tri-phonemes. We decided to annotate the datasets in terms of tri-phonemes so that the mouth position of the previous and following phonemes can also be considered in the synthesis process. This fact is important considering that phonemes can change their appearance depending on the previous or the following phonemes [146]. In Fig. 5.4 we illustrate the collecting process of the subject-specific generative dataset \mathcal{P}_G^s . Specifically, \mathcal{P}_G^s is defined as a superset that contains a collection of segments from all possible phonemes for each speaker. Afterward, we will consider the collected phonetic dataset \mathcal{P}_G^s together with additional text sentences to synthesize new video sequences that enrich our training set.

5.3.2.2 Collecting new grapheme transcriptions

Once the subject-specific phonetic datasets are collected, we need a new sentence corpus that covers a large vocabulary and includes word repetitions in varied contexts. In our case, we decided to take advantage of the open-source repository at *wikisource.org*, where there are many books freely available. We downloaded several books and preprocessed them to extract sentences with lengths between 3 and 12 words², similarly to the sentences recorded originally in the VLR dataset [61]. Then, we mapped the grapheme transcriptions into phoneme transcriptions using the linguistic rules detailed in [122] for the Spanish language and collected them in \mathcal{U}_G .

²The sentences are available at: <https://drive.google.com/file/d/1O68fJjTxKdbzZX5YfO5LGlJU2CbD2czK/view?usp=sharing>

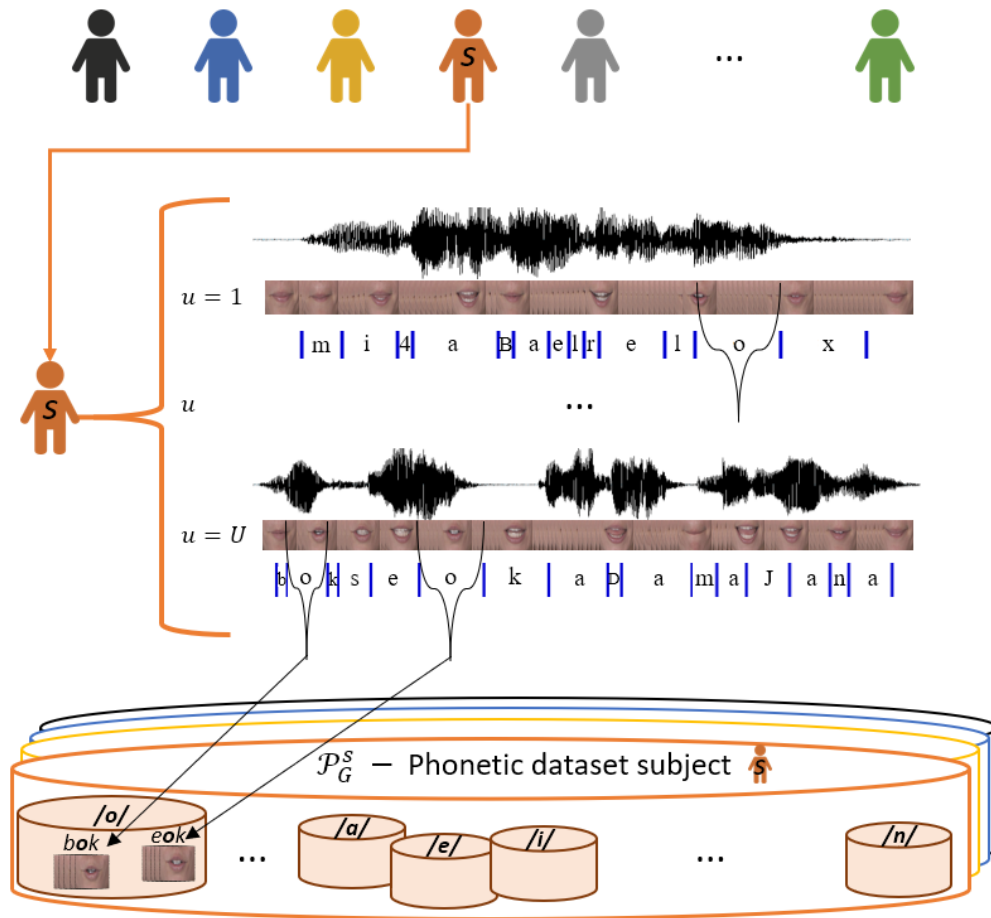


Figure 5.4: Collecting the phonetic dataset per subject

5.3.3 Synthesis of video sequences

Once the dataset of phonemes \mathcal{P}_G^s and the set of utterances \mathcal{U}_G have been defined, we follow the procedure shown in Algorithm 2, where we first define how many sentences N per speaker we want to generate from our set \mathcal{U}_G . Then, we follow an iterative process that selects the phoneme segments to form the new sentence u and interpolate them to generate a natural sequence. Finally, we verify the quality of the generated sequence and, if deemed acceptable, we save the generated video sequence. We explain these steps in detail in the following subsections.

5.3.3.1 Selection of phonetic segments

Firstly, a random sentence $u \in \mathcal{U}_G$ is selected. The synthesis procedure starts by picking one of the segments that correspond to the first phoneme uttered in

Algorithm 2 Synthesis of video sequences

```
Input  $\mathcal{U}_G, N, I$ 
for  $s = 1$  to  $S$  do
  Input  $\mathcal{P}_G^s$ 
  Initialize  $n = 0$ 
  while  $n < N$  do
    Pick a random  $u \in \mathcal{U}_G$ 
    Initialize  $Plausible = 0, iter = 0$ 
    while  $Plausible == 0$  and  $iter < I$  do
      Select the segments from  $\mathcal{P}_G^s$  to form the sentence  $u$ 
      Check the plausibility of the sequence
      Increment  $iter$  by 1
    if  $Plausible == 1$  then
      Save the generated sequence
      Remove  $u$  from  $\mathcal{U}_G$ 
      Increment  $n$  by 1
```

u and also the candidate to be the following segment that forms the sentence. However, considering that the mouth appearance for a given phoneme can change depending on its neighboring phonemes, the phonetic segments that follow after are not randomly selected. For each new segment, all candidate segments to produce the desired phoneme continuation are considered, analyzing each of them in descending order of merit until a plausible synthesis is achieved. If available, tri-phoneme combinations are prioritized, followed by two-phonemes and, finally, independent phonemes.

Each candidate segment is evaluated by appending its frames to the sequence synthesized so far and interpolating the transition³. Specifically, to account for noisy frame-based annotations, we establish a boundary window (W), proportional to the length of the phonetic segments, where we can find the optimal interpolation. Thus, for each pair of segments, the most plausible interpolation between them is the shortest window that is deemed plausible based on the criteria explained later in Section 5.3.3.2. If the interpolation is plausible, the evaluated segment is retained and appended to the synthesized sequence, as illustrated in Fig. 5.5. In contrast, in case the interpolation is considered implausible, we discard the current segment and repeat the evaluation process with the following candidate for the desired phoneme.

The above process continues until the sentence u is completed. The

³All the interpolations are performed using a state-of-the-art high-quality video frame interpolation method that is freely available in [155]

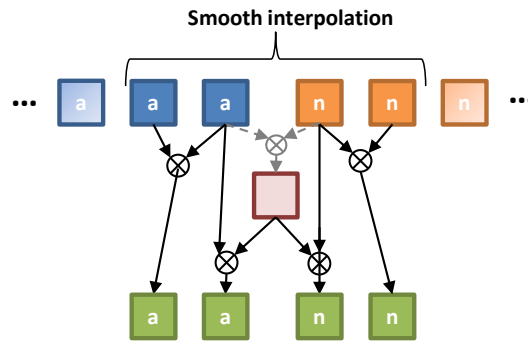


Figure 5.5: Interpolation example.

resulting video sequence can then be added to the training set, together with its corresponding grapheme annotation.

5.3.3.2 Video sequences plausibility

One requirement of visual speech synthesis is the plausibility of the generated sequences, in the sense that they look natural and that do not contain any visual artifacts. For example, in Fig. 5.6 we show some examples of non-plausible interpolations where the generated mouths contain many artifacts.

To avoid these artifacts and ensure natural video sequences, we should firstly analyze the statistics from real data, so that we can define what we consider a plausible sequence in a quantitative manner. To do so, we consider triplets of consecutive input frames $\{I_{i-1}, I_i, I_{i+1}\}$ and produce a synthetic estimate of the central frame by interpolation, which we denote \hat{I}_i . In this way, we can evaluate the quality of the interpolation by comparing the difference between the actual and simulated central frames. The rationale behind this, is that in a real video sequence, triplets of frames will correspond to smooth transitions and, thus, the interpolation process will successfully create a plausible transition between frames I_{i-1} and I_{i+1} . This, however, might not be the case when we try to merge phoneme segments taken from different utterances; e.g. the final mouth position from the



Figure 5.6: Wrong interpolation examples.

proceeding segment might not be sufficiently similar to the mouth position in the upcoming one, yielding an implausible transition.

To evaluate the quality of the interpolated frames, we assess the plausibility at both global and local levels:

1. *Global plausibility*: how likely is the interpolated frame to be generated from a linear model constructed from the frames of the same user under consideration;
2. *Local plausibility*: How similar is the color distribution of the interpolated frame with respect to the neighbouring frames.

Global plausibility is assessed by constructing user-specific PCA models; i.e. for each user, all its *real* frames are used to derive the principal components Φ^u , with eigenvalues Λ^u and mean μ^u . Thus, frame \hat{I}_t can be projected into PCA space as $\hat{\rho}_t = (\Phi^u)^T(\hat{I}_t - \mu^u)$. For every triplet $\{I_{i-1}, I_i, I_{i+1}\}$, we calculate the Mahalanobis distance and the reconstruction error of the simulated frame \hat{I}_i . The Mahalanobis distance, $d_m(i) = \hat{\rho}_i^T (\Lambda^u)^{-1} \hat{\rho}_i$ measures the likelihood of the synthesized frame once it is project to the PCA subspace, while the reconstruction error, $d_{pca}(i) = \|\hat{I}_t - \Phi^u \hat{\rho}_i - \mu^u\|$ indicates how far is \hat{I}_t from the PCA subspace.

Local plausibility is assessed by means of the Bhattacharyya distance. For every triplet $\{I_{i-1}, I_i, I_{i+1}\}$, we average the distances of the interpolated frame \hat{I}_t with respect to the previous and following frames, i.e. $d_b = (\rho(\hat{I}_t, I_{t-1}) + \rho(\hat{I}_t, I_{t+1}))/2$, where $\rho(\cdot)$ is the Bhattacharyya coefficient, defined in eq. (5.4).

Once the above metrics are computed for all triplets of consecutive input frames, we estimate plausibility thresholds for each of them based on standard outlier thresholds. Finally, the same metrics are also calculated every time an actual interpolation is computed to merge phoneme segments during the synthesis process. An interpolated frame is considered plausible if and only if all three metrics yield values below their corresponding outlier thresholds pre-computed in the training set. Recall that these thresholds are computed separately for each user and, as explained in the previous section, the exact position at which phoneme segments are merged is optimized within a boundary window (W), whose length is proportional to that of the segments to combine.

5.4 Proposed Lip-Reading Architecture

In this section, we introduce the architecture of the proposed ALR model. Figure 5.7 illustrates the system, which consists of two modules: the visual module, presented in yellow; and the temporal module, presented in blue. The convolutional or visual module receives color images of the mouth as input and

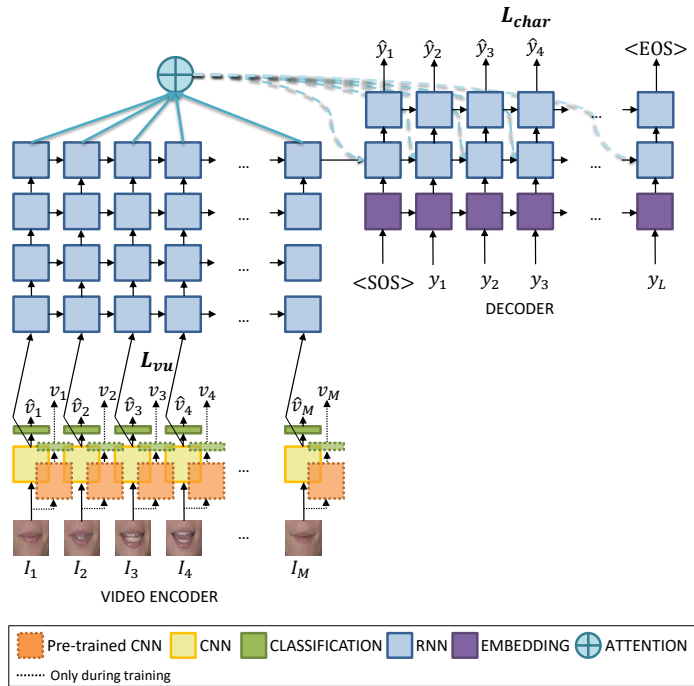


Figure 5.7: Proposed ALR system.

extracts feature vectors for each frame to encode the mouth appearance and also outputs a visual unit per frame. The sequential or temporal module is an attention-based seq2seq architecture, which consists of a video sequence encoder, a sequence decoder, and an attention mechanism. The encoder is based on stacked LSTMs that receive a sequence of relevant frame features and produces a latent representation (or memory) at each time step and a final latent state that summarizes the whole sequence. The decoder is also based on stacked LSTMs initialized with the encoder's final latent state. The decoder goal is to predict the expected speech units (i.e. characters or phonemes). The attention mechanism is added to help the decoder to cope effectively with long input sequences. Specifically, the attention mechanism is located between the encoder and the decoder to provide the decoder with information from each encoded hidden state. Thus, attention accesses the encoder memory at each time-step and provides a context vector that assists the decoder at each step. The attention model is able to selectively focus on useful parts of the input sequence and learn the temporal correspondence (or alignment) between them.

$$L = \frac{L_{vu}}{2} + \frac{L_{char}}{2} \quad (5.11)$$

The network minimizes a weighted sum of two losses (5.11), which consist of

Table 5.2: Visual module details

Name	Type	Filter Size \ Stride	Output Size
<i>conv0</i>	<i>Convolution</i>	$5 \times 5 \setminus 2$	$25 \times 25 \times 96$
<i>bn0</i>	<i>Batch Normalization</i>		$25 \times 25 \times 96$
<i>pool0</i>	<i>Max-Pooling</i>	2×2	$12 \times 12 \times 96$
<i>dp0</i>	<i>Dropout</i>	0.5	$12 \times 12 \times 96$
<i>conv1</i>	<i>Convolution</i>	$3 \times 3 \setminus 2$	$6 \times 6 \times 256$
<i>bn1</i>	<i>Batch Normalization</i>		$6 \times 6 \times 256$
<i>pool1</i>	<i>Max-Pooling</i>	2×2	$3 \times 3 \times 256$
<i>dp1</i>	<i>Dropout</i>	0.5	$3 \times 3 \times 256$
<i>conv2</i>	<i>Convolution</i>	$3 \times 3 \setminus 1$	$3 \times 3 \times 512$
<i>conv3</i>	<i>Convolution</i>	$3 \times 3 \setminus 1$	$3 \times 3 \times 512$
<i>conv4</i>	<i>Convolution</i>	$3 \times 3 \setminus 1$	$3 \times 3 \times 512$
<i>pool4</i>	<i>Max-Pooling</i>	2×2	$1 \times 1 \times 512$
<i>fc5</i>	<i>Fully-Connected</i>		512×1
<i>fc6</i>	<i>Fully-Connected</i>		512×1
<i>fc7</i>	<i>Fully-Connected</i>		$N \text{ classes} \times 1$

the cross-entropy loss L_{char} between the predicted character sequence $\hat{\mathbf{y}}$ and the real character sequence \mathbf{y} ; and the cross-entropy loss L_{vu} between the predicted sequence of visual units $\hat{\mathbf{v}}$ and an estimate of the visual units sequence \mathbf{v} assumed to be the ground truth (see Section 5.4.2).

We briefly describe each of these modules in the following subsections.

5.4.1 Input pre-processing

The database provides RGB images containing the whole face of the subjects. Therefore, we start detecting the face and performing an automatic location of the facial geometry (landmark location) using the Supervised Descent Method (SDM) [241]. Once the face is located, the estimated landmarks are used to fix a bounding box around the mouth region that is then normalized to a fixed size of 56×56 pixels. Taking into account that our visual module is based on VGG-M (input of 225×225) and that the latter has been trained with whole-face images, it seems reasonable to reduce the region considering the mouth size with respect to the whole face size.

5.4.2 Visual module

The visual front-end is based on VGG-M [33] and it processes random crops of 52×52 from the pre-processed images producing per-frame feature vectors of 512 dimensions and per-frame visual units \hat{v} . We present the model details in Table 5.2.

Notice that, as shown in Fig. 5.7, we build two CNN-based visual modules, one already pre-trained (shown in orange) and another one trained from scratch with the whole system (shown in yellow). This pre-trained network is crucial when planning to perform data augmentation techniques (e.g. synthesis of new video sequences). Even though we already generated a set of visual units \mathcal{V} that provides visual unit labels to our dataset, when new data is synthesized the pre-trained network allows to easily estimate its visual units. Let us emphasize that *pre-trained* here does not mean a network obtained from elsewhere, but one that our own visual module was trained on the classification task of visual units using only our original training set (i.e. real data, without augmentation). Once this is done, our ALR system can be trained from scratch and fully end-to-end, using the pre-trained visual network to provide *realistic* visual unit labels to any input frame. In this way, we are able to minimize the cross-entropy loss L_{vu} between the estimated visual units \hat{v} (i.e. outputs from the yellow CNN) and the ones estimated to serve as ground truth v (i.e. outputs from the orange CNN).

The pre-trained visual module uses the global set of visual units \mathcal{V} as training data. It is trained for 50 epochs using Adam optimizer with a learning rate of 0.0001 and mini-batches of 256 samples. The classifier is a fully-connected layer with a softmax that classifies among V visual units.

5.4.3 Temporal module

5.4.3.1 Encoder

The encoder consists of 4 stacked LSTMs with 1024 hidden units each. The first LSTM layer receives the visual features extracted by the visual module, and the final LSTM layer produces a latent representation (or memory) at each time step and a final latent state that summarizes the whole sequence.

5.4.3.2 Decoder

The decoder consists of two LSTM layers with 1024 hidden units and an attention layer. The attention’s goal is to generate a context vector c_l that captures useful information from the source-side to help the decoder to predict the correct output y_l . We used the global attention presented in [132], where the idea is to consider all encoder hidden states $\{h_1, \dots, h_m, \dots, h_M\}$ when deriving the context vector at

each decoding phase $l = \{1, \dots, L\}$. Specifically, at each decoding time step l , the model infers an alignment or weight vector α_l , whose size equals the number of time steps M , that scores the contribution of each encoder hidden state by comparing the current decoder hidden state h_l with each encoder hidden states h_m , as shown in (5.13). A global context vector c_l is then computed as the weighted average, according to α_l , over all the encoded states h_m , as shown in (5.14). The attention vector \mathbf{a}_l in (5.15) is then fed through a softmax layer to predict the output y_l .

The decoder predicts character sequences, where we can infer word-level results by splitting the decoding at blanks.

$$\beta(h_l, h_m) = h_l^\top \cdot W_a \cdot h_m \quad (5.12)$$

$$\alpha_{l,m} = \frac{e^{\beta(h_l, h_m)}}{\sum_{m'=1}^M e^{\beta(h_l, h_{m'})}} \quad (5.13)$$

$$c_l = \sum_m \alpha_{l,m} \cdot h_m \quad (5.14)$$

$$\mathbf{a}_l = \text{tanh}(W_c[c_l; h_l]) \quad (5.15)$$

The whole system is trained for 100 epochs using Adam optimizer with a learning rate of 0.0001 and mini-batches of 4 sequences. We trained the network using curriculum learning with a constant sampling of $\epsilon = 0.3$. Thus, during training sometimes the decoder input consisted of the real y_t instead of the predicted \hat{y}_t . To deal with over-fitting, we applied a dropout of 0.4 in both the encoder and decoder networks.

5.5 Experiments in a constrained scenario

In the literature, most ALR systems tackling with small-scale data are targeting recognition tasks in constrained scenarios such as word or sentence recognition, where the output of the system is limited to a pre-defined number of possible classes. Therefore, to be comparable with the state-of-the-art systems in small-scale scenarios, we propose, as a first step, to handle sentence-level classification without the need for large-scale databases. Specifically, we select the OuluVS2 database to evaluate our experiments as it is a widely used small-scale database.

Specifically, we proposed LDNet⁴ as a network that consists of a visual module based on VGG-M [33] and a temporal module that consists of a cascade

⁴Adapted from: Fernandez-Lopez, A., & Sukno, F. M. (2019, September). Lip-Reading with Limited-Data Network. In 2019 27th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE. DOI: 10.23919/EUSIPCO.2019.8902572.

of two LSTM layers with 256 hidden units that perform phrase classification at the end of the sequence, only after the whole stream has been processed.

5.5.1 Database

The OuluVS2 database [9] contains multi-view video recordings from 52 speakers uttering continuous digit sequences, short phrases and TIMIT sentences. We used the frontal-views of the second session where subjects were asked to read 10 daily-use English phrases. We tested our system in a speaker-independent setting. Following the testing procedure proposed by the database creators, we used 12 subjects for testing (s6, s8, s9, s15, s26, s30, s34, s43, s44, s49, s51 and s52) and 40 subjects for training and validation. Thus, we had 360 videos for testing (12 subj \times 10 phrases \times 3 repetitions), 1020 for training (40 subj \times 10 phrases \times 3 repetitions \times 0.85) and 180 for validation. We provide our results in terms of phrase-level classification (the standard for this dataset).

5.5.2 Results

In this case, instead of training the system fully end-to-end, we split the training by modules. In particular, we first pre-train the CNN to differentiate among weak visual labels, so that the features generated at the last step of the visual module (those at the FC layers) properly encode the mouth appearance and are helpful for the temporal module to decode visual speech. Once the visual module is trained, its classification layer is removed and the output from the FC layers is fed to the temporal module, which can be trained for phrase classification in a straightforward manner.

5.5.2.1 Visual module

We trained the visual module to classify among visually distinguishable units, which resulted in 13 visual units for the specific case of the OuluVS2 dataset. The CR obtained by the CNN module was 47.67%. While at first glance these results may seem modest, we will see that the features learned in this way by the CNNs are useful enough for the temporal module to produce high phrase recognition rates. Moreover, keeping in mind that our visual units are based on a similar definition to the one commonly used for visemes, our results are not far from those reported for phoneme and viseme classification in ALR [63, 218, 16, 113].

5.5.2.2 Temporal module

The temporal module shows the performance of the whole system because it outputs the spoken phrase. Following the procedure from [71, 177] we obtained

an average CR of 91.38% ($\pm 0.61\%$ standard deviation) averaged over 10 runs of temporal module training.

5.5.2.3 Comparison to other ALR systems

In this section we compare the DNN architectures evaluated in the OuluVS2 database (Table 5.3). Among systems using external training data, we firstly find the three systems proposed by Saitoh et al. [196]. Those systems used pre-trained models that were trained in external databases not related to lip-reading and were fine-tuned for OuluVS2. The GoogLeNet model achieved the maximum performance of 85.60% CR. Similarly, Chung and Zisserman [43, 44] proposed two systems specifically trained for lip-reading but pre-trained on much larger databases (LRW and LRS), and later fine-tuned for OuluVS2, achieving a maximum of 94.10% CR in [43].

There are several systems that do not use external data to train their model [117, 172, 176, 177, 71]. Among them, the most direct comparison to our system are those based on similar architectures, combining CNNs and LSTMs [117, 71]. The main difference between those systems and ours is the training process. In the case of Lee et al. [117], they directly trained their ALR system end-to-end from scratch, achieving a rather low performance of 81.10% CR. More recently, Fung and Muk [71] proposed a training strategy based on a big DA and on adding maxout activation units for ensuring better training. They achieved a higher accuracy (87.60% CR) with a system that combines 3D-CNNs with BiLSTMs. In contrast, in LDNet we follow a CNN-LSTM baseline, but propose an alternative training process. Specifically, we train the visual module separately to classify weakly labeled visual units, which are directly related to the spoken phrases. This has proven to be beneficial because it allows increasing the training samples while ensuring that the learned features are directly related to speech and not to other aspects such as speaker appearance. In this way, when the temporal module is added after the visual module, our system is able to achieve an average CR of 91.38%, which is quite competitive even with respect to systems using pre-trained models.

A different direction has been explored by Petridis et al. [172, 176, 177], who presented 3 systems based on an encoding network combined with BiLSTMs. Even though these systems do not follow the current trend in ALR, they reported 91.80% CR, which are state-of-the-art comparable results. However, analyzing these ALR systems we find that they were not trained end-to-end from scratch; instead, they pre-trained the encoding layer in a greedy layer-wise manner using Restricted Boltzmann Machines. They initialized their systems with the pre-trained encoder and trained the BiLSTMs while fine-tuning the encoder parameters. Compared to these 3 systems, LDNet provides a very similar

Table 5.3: Comparison with previous work on the OuluVS2 database.

With pre-trained models		Without pre-trained models	
Architecture	CR (%)	Architecture	CR (%)
CFI+NIN [196]	81.10	CNN+LSTM [117]	81.10
CFI+AlexNet [196]	82.80	Encoder+LSTM [172]	84.50
CFI+GoogLeNet [196]	85.60	Encoder+BiLSTM [176]	91.80
VGG-M+LSTM [43]	31.90	Encoder+BiLSTM [177]	91.80
SyncNet+LSTM [43]	94.10	CNN+BiLSTM [71]	87.60
CNN+LSTM+Att. [44]	91.10	LDNet(Ours)	91.38

accuracy, with low training time and maintaining a main-stream end-to-end ALR architecture, which is likely to benefit from the latest advances in the field, currently based on CNN-RNN architectures [62].

5.5.3 Conclusions

We investigate the design of an end-to-end ALR system that is simple to train without the need for large-scale databases. Specifically, we introduce an ALR system that performs phrase-level classification combining a visual module based on CNNs and a temporal module based on LSTMs. We show that thanks to the weak intermediate labels, it is feasible to obtain state-of-the-art performance by splitting the training by modules. We evaluated our system in the well-known OuluVS2 and reported a CR of 91.38% which is comparable to state-of-the-art results. Differently from previous approaches, our system does not require the use of any pre-trained model or external training data. LDNet training was completed in approximately 3.5 hours on a desktop computer with standard GPU hardware.

5.6 Experiments in continuous speech

5.6.1 The VLRf dataset

To evaluate the performance of the ALR system using a small-scale dataset, we selected the VLRf dataset [61] as one of the largest audio-visual databases in Spanish. Considering that most of the alternative languages to English suffer from insufficient data, this constitutes a relevant setting of wide applicability. We used the recordings of 24 speakers uttering 25 sentences each (from a pull of 500 sentences in total). The dataset was split using the 80/20 rule, where 20

sentences/user were selected for training (480 in total) and 5 sentences/user were reserved for testing (120 in total).

5.6.1.1 Data labeling

The VLRf database provides a phoneme label per frame. The phonemes used are based on the phonetic alphabet SAMPA [233]. For the Spanish language, the SAMPA vocabulary is composed of the following 31 phonemes: /o/, /m/, /k/, /w/, /t/, /j/, /l/, /x/, /L/, /u/, /g/, /z/, /d/, /G/, /4/, /r/, /T/, /b/, /j/, /s/, /e/, /p/, /n/, /N/, /J/, /B/, /D/, /i/, /tS/, /a/, /f/.

5.6.2 External language model

$$\mathbf{y}^* = \arg \max_y \frac{\log p(y|x) + \alpha \cdot \log p_{LM}(y)}{|y| - 1} \quad (5.16)$$

During inference we use an external character-level language model (LM) that consists of a RNN with 2 unidirectional LSTMs of 2048 units each. The LM is trained to predict one character at a time. Concretely, we trained the LM for 360 epochs with a learning rate of 0.001 and batches of 128 sequences using a set of $\sim 300k$ sentences⁵.

Decoding is performed using traditional beam search where the LM log-probabilities are combined with the model’s outputs via shallow fusion [4, 102]. Specifically, at inference time, we perform a log-linear interpolation at each time step (shown in (5.16)), where x represents the input sequence and y the output character sequence up to the current time step. The parameter α was fixed experimentally to 0.3.

5.6.3 Results

In this section we evaluate the classification rates of our system in terms of accuracy, Character Error Rate (CER) and Word Error Rate (WER).

5.6.3.1 Visual module

The capability of visual unit annotations To prove the validity of using visual units instead of phonemes to train the visual module, we compare the performance of our module when it is trained with visual units or with phonemes. Moreover,

⁵The sentences are available at: https://drive.google.com/drive/folders/1Xv9eQMfMv0yh9Cy415i3_vGnzB2NnNn?usp=sharing

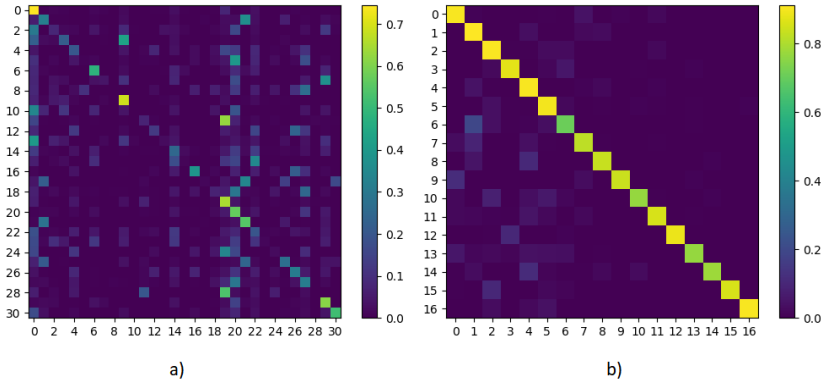


Figure 5.8: Confusion matrices: a) Phoneme labels from VLRf; b) Visual units derived using phoneme labels from VLRf. The phonemes can be found in the following order : /o/, /m/, /k/, /w/, /t/, /j/, /l/, /x/, /L/, /u/, /g/, /z/, /d/, /G/, /4/, /r/, /T/, /b/, /j/, /s/, /e/, /p/, /n/, /N/, /J/, /B/, /D/, /i/, /tS/, /a/, /f/.

we show how the performance when using visual units is not especially sensitive to the availability of phonetic labeling. To this end, we define the following experiments:

Exp. 1 : Train the visual module to classify *phonemes*. We use the labelling provided with the database [63] to perform a phoneme/frame classification, where the number of classes is fixed to $N=31$.

Exp. 2 : Train the visual module to classify *visual units*. We take advantage of the phoneme labelling to generalize our speaker-specific sets of visual units into a common set of visual units (details in Section 5.2.3). We obtain 17 visual units. Therefore, the number of classes is fixed to $N=17$ to perform a visual unit/frame classification.

In Fig. 5.8-a) we show the resulting confusion matrix of Experiment 1. In this case, we made use of the phoneme-level annotations per frame provided with the dataset. From that matrix, we can observe a high confusion between several phonemes, e.g. /m/ and /p/ (located at positions 1 and 21, respectively) or /s/ and /tS/ (located at positions 19 and 28, respectively), among others. The overall accuracy obtained over the 31 phonemes is 46.24%. This result is not very high, which is not surprising since the literature has shown that there is no one-to-one mapping between acoustic and visual information [18, 87, 27, 7].

When we derive visual units that, while imperfect, are generated automatically in such a way that they consider speech but also the visual similarity between mouth positions, we would expect to observe a clear improvement. Specifically,

we should end up with more discriminative features that allow disambiguation among different lip positions under the same phoneme. This effect can be easily observed in Fig. 5.8-b, where we show the resulting confusion matrix among the 17 obtained visual units, obtaining a CR of 85.25%. Those 17 visual units were obtained by the procedure described in Section 5.1.1 for the sentences of the VLRf dataset.

In both experiments, we performed the same 80/20 training/test split indicated in Section 5.6.1 where all frames corresponding to the 120 test sentences were used to compute the confusion matrices and the rest (480 sentences) were used for training. Therefore, even though the accuracy tends to grow as we reduce the number of classes, the comparison between both confusion matrices (Fig. 5.8-a&b) clearly shows the effectiveness of the visual module when trained with visual units instead of phonemes.

In the first two experiments above, we made use of the labeling of all input frames in terms of phonemes provided with the VLRf dataset, either as the target to train the visual module (in Experiment 1) or to guide the derivation of a common set of visual units from speaker-specific units, as described in Section 5.2.3. However, this implies an undesirable manual pre-processing load, which is time-consuming and non-feasible in many cases. Fortunately, we will show in the following experiments that the effectiveness of the visual module is not especially sensitive to a very accurate phoneme labeling. As a consequence, **it is possible to derive visual units from approximate and fully automatic phonetic annotations**. Specifically, as detailed in Section 5.2.3, we use a rough estimate of the phoneme label for each frame based on the mean duration of phoneme, leading to the next two experiments:

Exp. 3 : Train the visual module to classify *approximated phonemes*. We use the mean phoneme duration to estimate a labelling and perform a phoneme/frame classification, where the number of classes is fixed to $N=31$.

Exp. 4 : Train the visual module to classify *visual units*. We repeat the procedure from 5.2.3) to generalize our speaker-specific sets of visual units into a common set of visual units, but using the approximated phoneme labelling instead of the manual labeling provided with the database. We obtain 25 visual units. Therefore, the number of classes is fixed to $N=25$ to perform a visual unit/frame classification.

Fig. 5.9-c) shows the resulting confusion matrix using the 31 phonemes under the approximate labeling. In this case, the CR is very low, just above 13%. This fact makes sense considering that the boundaries between phonemes

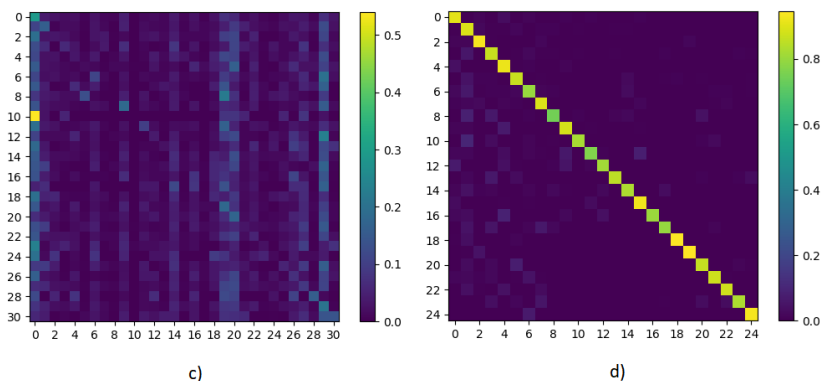


Figure 5.9: Confusion matrices: c) Estimated phoneme labels; d) Visual units derived using estimated phoneme labels. The phonemes can be found in the following order :/o/, /m/, /k/, /w/, /t/, /j/, /l/, /x/, /L/, /u/, /g/, /z/, /d/, /G/, /4/, /r/, /T/, /b/, /j/, /s/, /e/, /p/, /n/, /N/, /J/, /B/, /D/, /i/, /tS/, /a/, /f/.

become much noisier. Interestingly, even though these approximated annotations are worse compared to the manually-labeled phonemes, they are good enough to approximate the phonetic distributions of speaker-dependent visual units. This can be seen in Fig. 5.9-d), which shows the performance using the visual units derived from the approximate phoneme labeling. Indeed, the obtained results are even better than those from Experiment 2, even when the number of resulting classes is higher, yielding a CR of 87.70%.

5.6.4 Temporal module

5.6.4.1 Architecture details

The ALR system is trained end-to-end for 100 epochs using Adam optimizer, batches of 4 and a learning rate of 0.0001. The classifiers are two softmax layers that use cross-entropy loss to classify among 25 visual units and 28 characters (including space). We used beam search decoding with an external language model (Sec. 5.6.2) with $B = 20$ at inference time.

Continuous speech decoding In this section, we compare how much we can lipread from the VLRf dataset under different setups and ALR systems.

Exp. 5 - Baseline: Train an ALR system from scratch to perform character-level speech recognition. Consider only the 480 sentences for training available in the VLRf dataset. Minimize the cross-entropy loss to classify among 28 characters (including space).

Table 5.4: Character Error Rate (CER) and Word Error Rate (WER) for ALR systems \mathcal{A} and \mathcal{B} evaluated on the VLRf dataset for different experimental conditions. We show our results when using Greedy (G) decoding, Beam Search (BS) decoding and an additional external language model (LM) with BS.

Experiment	Architecture	G – CER/WER	BS – CER/WER	BS&LM – CER/WER
Exp. 5	\mathcal{A}	81.18 / 106.81	76.80 / 97.76	74.90 / 97.41
Exp. 6	\mathcal{A}	69.14 / 93.62	69.75 / 93.92	70.50 / 93.59
Exp. 7	\mathcal{A}	78.56 / 97.98	79.55 / 102.60	78.18 / 97.48
Exp. 8	\mathcal{A}	49.92 / 91.52	46.78 / 87.14	51.87 / 77.80
Exp. 5	\mathcal{B}	81.61 / 114.79	77.31 / 107.98	79.06 / 101.36
Exp. 6	\mathcal{B}	74.40 / 100.00	76.44 / 99.51	75.17 / 98.26
Exp. 7	\mathcal{B}	78.98 / 95.93	79.88 / 105.24	84.09 / 108.84
Exp. 8	\mathcal{B}	48.61 / 89.57	44.77 / 82.42	48.81 / 72.90

*Exp. 6 - **Baseline + VU***: Train an ALR system from scratch to perform character-level speech recognition. Consider only the 480 sentences for training available in the VLRf dataset. Minimize the cross-entropy loss at the end of the system to classify among 28 characters (including space) and the cross-entropy loss at the end of the visual module to classify among 25 visual units (VU).

*Exp. 7 - **Baseline + DAS***: Train an ALR system from scratch to perform character-level speech recognition. Consider the 480 sentences for training available in the VLRf dataset plus the Data Augmentation by Synthesis (DAS), consisting of $\sim 50,000$ video sequences. Minimize the cross-entropy loss to classify among 28 characters (including space).

*Exp. 8 - **Baseline + DAS + VU***: Train an ALR system from scratch to perform character-level speech recognition. Consider the 480 sentences for training available in the VLRf dataset plus the synthesized video sequences ($\sim 50,000$). Minimize the cross-entropy loss at the end of the system to classify among 28 characters (including space) and the cross-entropy loss at the end of the visual module to classify among 25 visual units.

Table 5.4 summarizes the CER and WER results for Experiments 5-8. Specifically, we repeated the above 4 experiments for 2 different network setups, which we denote networks \mathcal{A} and \mathcal{B} . Network \mathcal{A} consists of 3 encoding layers followed by 2 decoding layers. All of them with 512 hidden units. In contrast, network \mathcal{B} contains 4 encoding layers and 2 decoding layers with 1024 hidden units each. The two different architectures are reasonably standard for current state-of-the-art ALR systems; their hyper-parameters have been chosen with the

aim of achieving a good balance between the available data and the number of parameters of the network.

From Table 5.4, we can observe that when we train our ALR system using only the 480 sentences available from the VLR dataset (Exp. 5), the results are not satisfactory, and we may say that we are not really able to lip-read, independently of whether we use network \mathcal{A} or \mathcal{B} . This is an expected consequence of the lack of sufficient data to train powerful DNNs and highlights the need for alternative training strategies, as discussed in Sec. 5.

To tackle this problem, we introduce visual units to the training procedure in Experiment 6, as detailed in Section 5.4. Experiment 6 improves between 3% to 5% of CER and WER with respect to Experiment 5, in both \mathcal{A} and \mathcal{B} setups. We can also observe that the smaller DNN tends to perform better in this case, as expected given the use of a small-scale training set. However, these results are still far from the state-of-the-art in lip-reading ($\sim 35\%$ CER and 50% WER on large scale datasets [4]), where the success of those ALR systems relies on the interpretability of the context.

Therefore, in the subsequent experiments, we train our system with a much larger training set by using new synthesized video sequences as a means for data augmentation (Section 5.3). Indeed, in Experiments 7 and 8 we analyze if the only requirement to adequately train our ALR system is simply to have more data or if the knowledge added by the visual units is also helpful. Then, Experiment 7 was trained with more than 50k video sequences (i.e. real and synthesized data) to perform character-level classification. However, we see that the system does not improve with respect to the previous experiment in any of the setups. This suggests that the visual features are not being properly learned to help the temporal module to predict the sequence. In contrast, Experiment 8 shows that when the same setup incorporates the visual units' loss, results improve dramatically and the resulting system is able to lip-read. In particular, we can observe that the best results are achieved for model \mathcal{B} , which consists of 4 encoding and 2 decoding layers, obtaining a 48.8% of CER and 72.9% of WER. These results represent a significant improvement of 34% in CER and 28% in WER with respect to the baseline in Experiment 5.

5.6.4.2 Attention visualization:

The attention mechanism between the encoder/decoder model generates alignment between the video sequence and the expected character output. Figure 5.10 reproduces a visualization of the attention alignments of the sentence *ayer me tope con un chico muy guapo* where the colors indicate the weight of the attention mechanism between the video frames in the video source and its corresponding transcription in characters. This visualization demonstrates that the ALR model

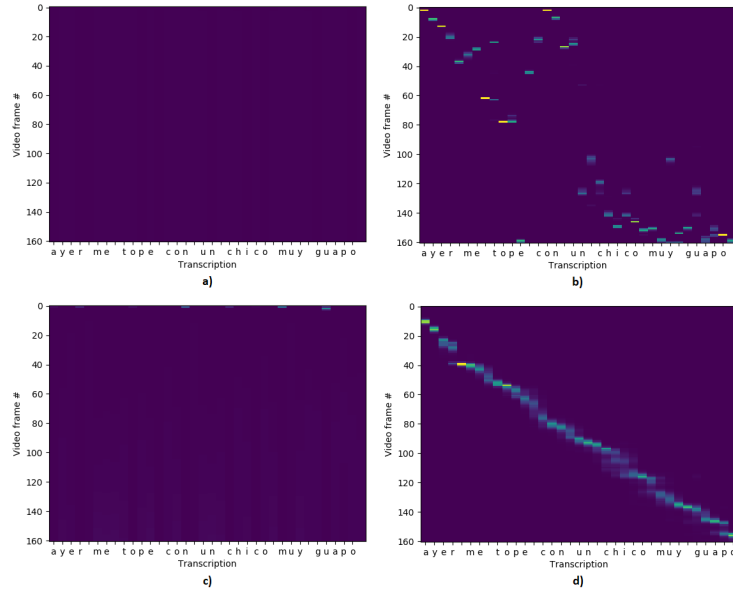


Figure 5.10: A visualization of attention weights evaluated on system \mathcal{B} : a) Baseline; b) Baseline + VU; c) Baseline + DAS; d) Baseline + DAS + VU. We only observe soft-alignment between source video sequence and target characters for Baseline + DAS + VU.

has successfully learned a soft alignment between frames and output characters.

5.6.4.3 Decoding examples:

Table 5.5 shows some examples of sentence predictions given by the ALR model.

5.6.5 Discussion and Conclusions

In this work, we explore continuous visual speech recognition at the character-level when the available data is limited. In particular, we selected the VLR database, which is one of the largest datasets in Spanish but is still more than 100 times smaller than the largest English-spoken one, a problem that is widespread for non-English languages. In spite of this, we reached a performance that is competitive with the state-of-the-art systems presented in English [11]; [41]; [202]; [134], thanks to our proposal of an alternative learning strategy that allows end-to-end training of an ALR system without the need for large-scale data when the proper restrictions are introduced to the visual module. Specifically, we introduced a self-supervised training strategy that takes advantage of intermediate labels, termed *visual units*. These visual units, defined as *a collection of visually*

Table 5.5: Examples of ALR results. **GT**: Ground Truth; **G**: Greedy, **BS**: Beam Search; and **LM**: Language Model.

	Transcription	CER/WER
GT	se oyeron los disparos de los cazadores	
G	se oieron los simagos y de los adorissr	33.3 / 50.0
BS&LM	se oyeron los disparos de los adoradores	10.0 / 14.2
GT	mi barrio se llama canaletas	
G	pi maricosse ama a aaaaaess	48.1 / 100.0
BS&LM	mi marido se llama carañas	26.9 / 40.0
GT	ayer me tope con un chico muy guapo	
G	ayer me to pe consus icoommulugaapo	28.5 / 100.0
BS&LM	ayer me tome con un disco muy guapo	11.4 / 25.0

similar images constrained by linguistics, are informative enough about the mouth and lips position and are generated in a fully-automatically manner without any human intervention.

In this way, we show that a CNN-based model is able to differentiate among visually separable classes, i.e. single-frames representing different mouth positions related to speech, and that the features generated at the last layer of the visual module properly encode the mouth appearance and really help the temporal module to decode visual speech and predict the correct character. Consequently, our ALR system was trained simultaneously for the multitasking classification of visual units and characters. Additionally, we also tackled the issue of reduced linguistic content by implementing spatiotemporal data augmentation. We took advantage of a phonetic annotated dataset and generated new video sequences by appropriately combining character-like sub-sequences from existing ones.

We evaluated our ALR system using the VLRFB dataset, a phonetically annotated dataset in Spanish. It is important to highlight that the VLRFB database is one of the largest audio-visual datasets in Spanish and consists of 600 sentences in total. Therefore, from the only 480 sentences for training, we were able to perform data augmentation and synthesize $\sim 50,000$ new sentences to train the whole system end-to-end. In our experiments, we showed that the ALR system requires both the visual units and the synthesized sentences to be properly trained, but that their separate use would not be satisfactory. We achieved a 44.77% CER and 72.90% WER on the test data, for an ALR system that was trained with less than 500 utterances in less than 2 days on a single Tesla GPU.

Even though the above results may seem still modest when compared to the top-results reported in English-spoken data, such comparison must be addressed

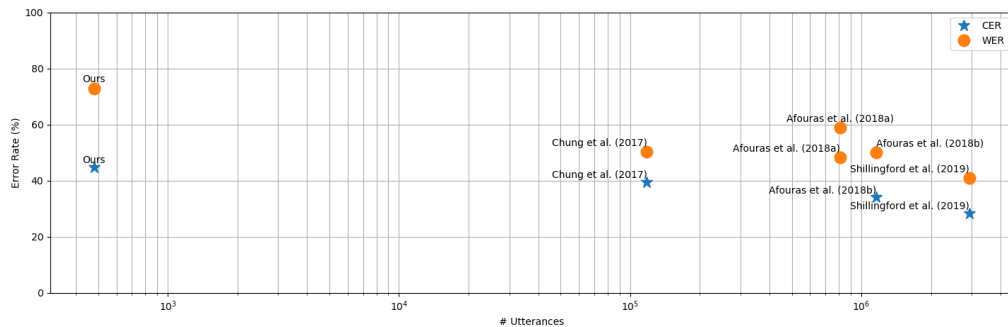


Figure 5.11: CER and WER for different ALR systems conditioned to the number of training sentences.

considering the amount of available data. To this end, Fig. 5.11 shows the CER and WER obtained by different state-of-the-art ALR systems and the amount of data that each of them used for training. It can be seen that the higher the available training data, the lower the error rates. For this reason, most of the systems displayed in this figure were trained on a combination of multiple datasets to increase the number of training samples. Thus, when comparing those systems to ours, the amount of training data differs by more than 100 times. Concretely, the best performance to date was reported by [202] using a training set of approximately 3 million sentences; comparatively we used only about 0.024% of their data. In this sense, our results are arguably the best to date for this volume of training material. Moreover, our method allows training an end-to-end system in less than 2 days on a single GPU, while most of the state-of-the-art architectures require between 8 and 13 days and considerably more computation power. These results open the door to continuous lip-reading in many different languages, for which small-scale datasets already exist or can be generated without excessive effort. Future research will focus on transfer learning between languages, e.g. some visual units can be common to a large number of subjects, independently of their language.

Chapter 6

VISUAL SPEECH ADAPTATION TO NEW SPEAKERS

Speech perception is inherently a multi-modal phenomenon that involves both acoustic and visual cues. This has led to research in Audio-Visual Speech Recognition (AVSR) systems, which try to balance the contribution of the audio and the visual information channels to develop systems that are robust to audio artifacts and noise. However, exploiting both modalities simultaneously has proven challenging. Firstly, current systems still rely strongly on the audio stream. Secondly, although visual information is not affected by acoustic noise, it is well known that the access to speech through the visual channel is subject to several limitations [4, 41]. For instance, the fact that every person pronounces in a unique way makes the generalization of visual models very difficult [119]. Indeed, it has been shown that lip-reading accuracy is more consistent in Speaker Dependent (SD) scenarios [243, 220, 17, 228, 229, 16, 230, 211, 30], and moreover, that human lip-reading accuracy increases when a relationship is developed between the speaker and the lip-reader [123], i.e. a lip-reader can learn to read one individual more accurately than to lip-read the general population [30, 61]. However, for AVSR systems to operate in realistic settings there is the need to target natural speech of any speaker. Although features are highly SD [47, 7], SD systems have a limited applicability due to their requirement for large amounts of data from the target speaker. Accordingly, there is a lot of interest in speaker adaptation techniques because they can be combined with SI systems and exploit speaker-dependencies to improve the recognition rates. Visual adaptation offers

Adapted from: Fernandez-Lopez, A., Karaali, A., Harte, N., & Sukno, F. M. (2020, May). Cogans For Unsupervised Visual Speech Adaptation To New Speakers. *In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6294-6298). IEEE. DOI: 10.1109/ICASSP40776.2020.9053299

advantages over audio-only adaptation as the video is not affected by noise in the available audio data for a new speaker, which may also be limited.

Contribution: in this work we focus on the visual adaptation of an SI-AVSR system to an unknown and unlabeled speaker. In particular, we adapt an AVSR system trained in a source domain to decode samples in a target domain without the need for labels in the target domain. Our system jointly addresses the speech recognition problem and the unsupervised speaker adaptation problem. For Unsupervised Domain Adaptation (UDA) we propose the use of Coupled Generative Adversarial Networks (CoGAN) [121], where a joint distribution of multi-domain images is learned without the existence of corresponding images between different domains. This is the first time that CoGANs are adopted for audiovisual speaker adaptation that the authors are aware of. We evaluate our character-based AVSR system on the large scale audio-visual TCD-TIMIT dataset. Our results show an average improvement above a 10% with respect to its audio and audio-visual equivalents.

6.1 Speaker Adaptation

As visual features are highly speaker-dependent, we hypothesize that visual speech adaptation will increase the contribution of the visual channel, improving the overall performance of an AVSR system, especially in noisy scenarios. For that reason, we investigate the UDA of the visual front-end of an SI-AVSR system to an unknown and unlabeled speaker. UDA adapts a model trained in a source domain to generalize on samples from a target domain where there is no annotated data for re-training.

We define X^1 as the data to train the SI-AVSR system. Each sample of X^1 comprises a variable length acoustic sequence $\{a_1^1, a_2^1, \dots, a_N^1\}$, its corresponding visual track $\{v_1^1, v_2^1, \dots, v_M^1\}$ and its grapheme transcription $\{y_1^1, y_2^1, \dots, y_L^1\}$. All those samples belong to a set of multiple sequences $p \in [1, P^1]$ of different speakers ($s_i \in S^1, i \in [1, K]$). In contrast, the second domain X^2 concerns a small set of sentences $p \in [1, P^2]$ of only the visual track $\{v_1^2, v_2^2, \dots, v_M^2\}$ of a single unknown speaker $s_j \in S^2, j \notin [1, K]$, where no character-labels are available.

Following the above, we propose to jointly tackle: i) the speech recognition problem, addressing the construction of a robust AVSR system; and ii) the speaker adaptation problem, to adapt an SI system to an unseen and unlabeled speaker. For the speech recognition problem we use all data from X^1 , consisting of a set of audio-visual sequences with annotations. In contrast, for the speaker adaptation problem we use video-frames v from both domains X^1 and X^2 .

We assume that in an ideal SI speech recognition system, the same speech

events should be encoded in a similar way, independently of the speaker. Therefore, when processing visual cues, the visual front-end should encode equivalent lip positions from different speakers with a similar representation. Nevertheless, video-frames v^1 and v^2 are samples from two different marginal distributions p_{X^1} and p_{X^2} for which tuples of corresponding images are not available (i.e. we have no correspondence between lip positions). Thus, we propose to use a CoGANs framework to learn a joint distribution of multi-domain images in an unsupervised manner. Once the joint distribution is learnt, we can automatically adapt the system to an unseen speaker by minimizing the distance between the features generated by the lip images that were identified to correspond each other across the two domains.

6.2 Proposed AVSR System

We aim to adapt an SI-AVSR system to an unknown and unlabeled speaker. To this end, we combine an AVSR system based on the AV Align system recently presented in [207] with a domain adaptation network based on CoGANs.

6.2.1 CoGANs

CoGANs [121] address the problem of learning a joint distribution of multi-domain images from data without the existence of corresponding images from the different domains. The only requirement is a set of images drawn from the marginal distributions of each domain. CoGANs consist of a pair of GANs, each one responsible for synthesizing images in one domain. Their novelty is that by enforcing a weight-sharing constraint between both GANs, the networks learn to synthesize corresponding images without correspondence supervision. Based on the idea that deep neural networks learn hierarchical representations, both GANs are forced to share the same high-level concepts. In the case of the generative models, which gradually decode information from more abstract concepts to more specific details, the first layers share the same parameters. In this way, we force both models to generate images that share the same semantics (e.g. same lip position) but different details (e.g. different skin color, lip width or other details that are more related to identity than speech). In contrast, for the discriminative models, where the flow is completely opposite, the last layers are the ones that are shared because those are the ones responsible for extracting higher-level features.

Let v^1 and v^2 be lip-images from distributions p_{X^1} and p_{X^2} . We define a couple of GANs that consists of two generative models G^1 and G^2 and two discriminative models D^1 and D^2 . G^1 and G^2 map from a common random normal vector $z \in \mathbb{R}^d$ into two images with similar distributions \bar{v}^1 and \bar{v}^2 , where those images are

corresponding, i.e. they show the same lip-position for different speakers ($s_i \in S^1$ and $s_j \in S^2$). To ensure this lip-correspondence, G^1 and G^2 share the high-level semantics corresponding to the parameters θ of all layers except the last one ($\theta_{G^1_{(l)}} = \theta_{G^2_{(l)}}$, for $l = 1, 2, \dots, L_G - 1$), which will be responsible for encoding the details of each specific domain X^1 and X^2 . In contrast, D^1 and D^2 map an input image to a probability score that determines if a sample belongs to the real data distribution. Those models are CNN-based, where the last layers represent the higher-level features. Therefore, the parameters θ of all the layers except the first one are shared ($\theta_{D^1_{(l)}} = \theta_{D^2_{(l)}}$, for $l = 2, 3, \dots, L_D$), i.e. L_G and L_D are the number of layers in $G^{1,2}$ and $D^{1,2}$. CoGANs aim to model the joint distribution following:

$$\min_{G^1, G^2, D^1, D^2} \max \mathcal{L}_{GAN}(G^1, G^2, D^1, D^2) \quad (6.1)$$

where the objective function is given by (6.2) and subject to (6.3):

$$\begin{aligned} & \mathbb{E}_{v^1 \in p_{X^1}}[\log(D^1(v^1))] + \mathbb{E}_{z \in p_z}[\log(1 - D^1(G^1(z)))] + \\ & \mathbb{E}_{v^2 \in p_{X^2}}[\log(D^2(v^2))] + \mathbb{E}_{z \in p_z}[\log(1 - D^2(G^2(z)))] \end{aligned} \quad (6.2)$$

$$\begin{aligned} \theta_{G^1_{(l)}} &= \theta_{G^2_{(l)}}, \text{ for } l = 1, \dots, L_G - 1 \\ \theta_{D^1_{(l)}} &= \theta_{D^2_{(l)}}, \text{ for } l = 2, \dots, L_D \end{aligned} \quad (6.3)$$

6.2.2 Proposed architecture

We used the AV Align system [207, 208] as baseline. The system code is publicly available and consists of an attention-based sequence-to-sequence (Seq2seq) model that fuses audio and visual information in the feature space and has been evaluated with datasets such as TCD-TIMIT and LRS2. Audio features are based on log Mel features while visual features are obtained by processing the images through a ResNet-CNN. Interestingly, the Seq2seq network fuses the audio-visual information at the encoder side (instead of the decoder side as done in most AVSR systems [42]) by explicitly aligning the acoustic features with the visual ones in an unsupervised manner using a cross-modal attention mechanism. This fact is particularly important because it helps the Seq2seq model to pay more attention to the visual information (i.e. the acoustic encoder’s top layer can no longer be considered as an acoustic-only representation).

In Fig. 6.1 we present our Adapted-AVSR system. The network consists of the feature extraction blocks for audio and visual cues, two sequence encoders, one for each modality, a joint audio-visual sequence decoder, and two attention

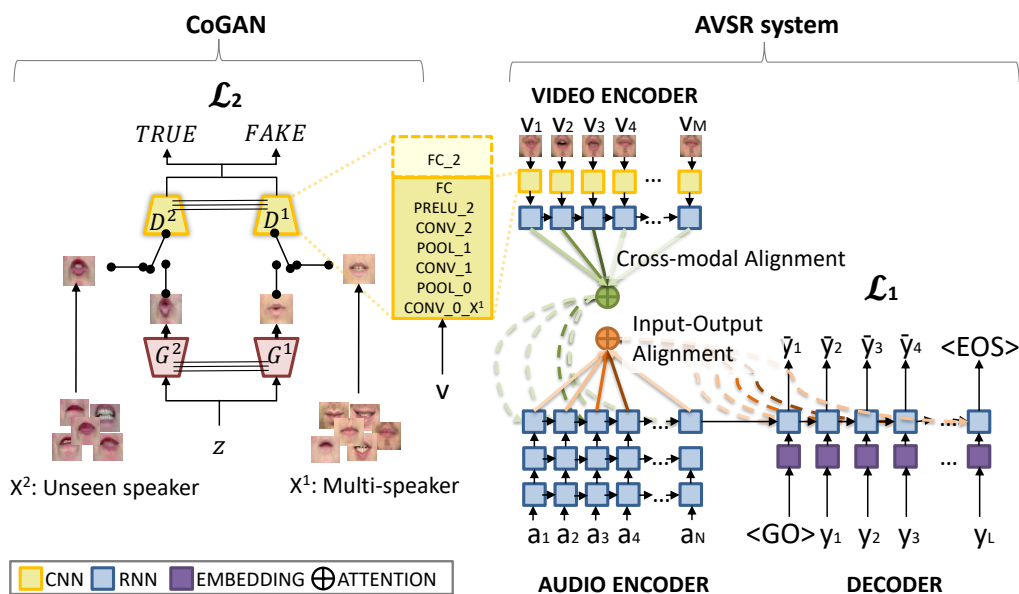


Figure 6.1: Proposed Speaker Adapted-AVSR system. On the left we see the feature adaptation using CoGANs: On the right, the AVSR system. Both networks are jointly trained to adapt the visual front-end to a new speaker.

mechanisms, one for the cross audio-visual alignment and another one for the input/output alignment. For the audio features, we follow the same transformation process in log Mel features as [207]. In contrast, the visual front-end consists of a CNN, which processes images of 36×36 pixels and generates a vector of size 128. The audio encoder consists of 3 LSTM layers with 256 hidden units, while the video encoder consists of a single LSTM layer with 256 hidden units. We used the cross-modal alignment architecture [207], where the acoustic representations are explicitly aligned with the visual ones in an unsupervised way. Finally, a character-based LSTM decoder with 256 hidden units attends the enhanced audio-visual representations. In Table 6.1 we present the details of the CoGAN architecture. The CoGANs and the AVSR system are connected by sharing the same parameters between the discriminator D^1 and the visual front-end.

Our AVSR system minimizes \mathcal{L}_1 , the cross-entropy loss (6.4) between the real character sequence $Y = \{y_1, y_2, \dots, y_L\}$ and the decoded character sequence $\bar{Y} = \{\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L\}$. In contrast, the CoGAN aims to adapt the model to an unseen speaker by: i) learning a joint distribution between lip-images; and ii) reducing the distance between the feature extraction of those corresponding lip-images. As shown in the last term of (6.6), we minimize the distance of the penultimate convolutional layer for both GANs because they correspond to the extracted

features (i.e. last layer of the visual front-end). Thus, we minimize the distance between the extracted features of corresponding images in both domains. We are able to adapt the SI-AVSR to the unknown speaker by iteratively minimizing equations (6.4) and (6.5) subject to (6.3).

$$\min_{\bar{Y}} \mathbb{E}_Y[-\log \bar{Y}] \quad (6.4)$$

$$\min_{G^1, G^2, D^1, D^2} \max \mathcal{L}_2(G^1, G^2, D^1, D^2) \quad (6.5)$$

$$\mathcal{L}_2 = \mathcal{L}_{GAN} + \lambda_1 \| D_{(L_D-1)}^1(G^1(z)) - D_{(L_D-1)}^2(G^2(z)) \| \quad (6.6)$$

Table 6.1: GANs architecture for speaker adaptation

Generator			Discriminator		
Operation	Parameters	θ	Operation	Parameters	θ
TConv+BN+PReLU	F=1024, K=4, S=1	Y	Conv+Pool	F=20, K=7, S=1	N
TConv+BN+PReLU	F=512, K=3, S=2	Y	Conv+Pool	F=50, K=7, S=1	Y
TConv+BN+PReLU	F=256, K=3, S=2	Y	Conv+PReLU	F=500, K=4, S=1	Y
TConv+BN+PReLU	F=128, K=3, S=2	Y	FC	F=128, K=1, S=1	Y
TConv+Sigmoid	F=3, K=6, S=1	N	FC	F=2, K=1, S=1	Y

6.3 Experiments and Results

6.3.1 The TCD-TIMIT dataset

To evaluate the adaptation of our AVSR system to a new speaker we need a dataset in which the set of sentences per speaker is big enough to train the CoGANs and to evaluate complete system performance. Among the available datasets, the TCD-TIMIT is the largest one that fulfills this requirement [62]. The TCD-TIMIT dataset [86] is a large-scale multi-speaker audio-visual database in English. The audio-visual data contains recordings of 59 speakers (32 male and 27 female) uttering 96 sentences selected from a pool of 6,913 of phonetically balanced (*sx*) and diverse (*si*) sentences. Specifically, there are 450 *sx* sentences spoken by 7 different speakers and 36 *si* sentences that are unique to each speaker. The dataset has been split into training/test aiming to balance gender and facial hair. Therefore, speakers 06M, 14M, 17F, 18M, 31F, 41M, 46F, 47M, and 51F are selected for the test, and the rest for training.

6.3.2 Training procedure

We train audio-only (ASR), audio-visual (AVSR) and speaker adapted audio-visual (*Adapted-AVSR*) systems to decode continuous speech at the character level. We additively corrupt the acoustic modality with *Cafeteria Noise* (shown to be the most challenging in [207]) and expose our systems to four different levels of noise, firstly clean speech and then with a Signal-to-Noise Ratio (SNR) of 10db, 0db, and -5db. The ASR and AVSR systems minimize the cross-entropy loss between Y and \bar{Y} . In contrast, the *Adapted-AVSR* system iteratively minimizes (6.4) and (6.5), as explained in section 6.2. λ_1 from (6.6) was settled experimentally to 0.01.

Maintaining the same training/test procedure for all systems, we used $P^1=4800$ sentences (50×96) for training and 36 *si* sentences per speaker for test (36×9). We evaluated the system using only the unique *si* sentences per speaker to properly analyze the influence of visual information, where systems have no previous knowledge of the sentences. The video-images v of the remaining $P^2=60$ *sx* sentences per speaker were used to train the Adapted-AVSR system. This yields 9 Adapted-AVSR systems, one for each test speaker.

6.3.3 Learning a joint distribution

The most important factor in learning a joint distribution is the weight-sharing constraint between the generative models. The first layers encode the same mouth position, while the last layer encodes the speaker's details. However, the effect of weight-sharing between the discriminative models is unclear. In [121], they showed that CoGANs learn a joint distribution without weight-sharing layers in the discriminative models. In Section 6.2-eq.(6.6), we explain that the speaker adaptation involves minimizing the distance between the feature vectors of correspondent images in both domains. This suggests that sharing more layers facilitates the minimization of the distance between the features. Therefore, we look for the maximum number of shared layers to ensure the learning of the joint distribution. In Fig. 6.2-(a) we can observe the generated images from G^1 and G^2 for $D^1 \equiv D^2$, i.e. all layers are shared. Here, for each random vector z , the generative models produce the same lip position but for details of the same speaker, i.e. D^1 and D^2 , cannot distinguish between domains X^1 and X^2 and the generative models beat them by converging to the same point. In contrast, in Fig. 6.2-(b), we show that not sharing the first layer between D^1 and D^2 is enough to ensure the learning of the expected joint distribution.

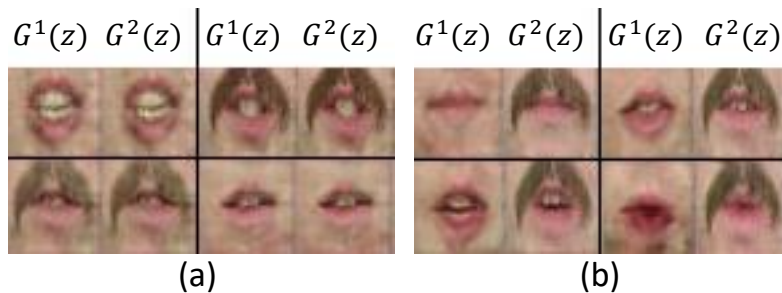


Figure 6.2: Synthesized images for different z when: (a) $D^1 \equiv D^2$; (b) $\theta_{D^1_{(l)}} \neq \theta_{D^2_{(l)}}$ and $\theta_{D^1_{(l)}} = \theta_{D^2_{(l)}}$, for $l = 2, 3, \dots, L_D$

6.3.4 Comparison between speech recognition systems

In Fig. 6.3-(a) we show the average Character Error Rate (CER) across the systems evaluated on the 9 different speakers. The visual adaptation improves system accuracy, outperforming the other systems in all scenarios. Specifically, we observe an improvement between 8%-15.5% with respect to the ASR system and 6%-10% with respect to the AVSR system. It is important to highlight that our CER, while comparatively lower than that achievable on larger datasets, is close to the performance achieved in [207] with speaker-dependent experiments.

Counter-intuitively, the performance improvement of AVSR against ASR reduces as the audio becomes noisier. This occurs because the cross-modal alignment deteriorates when the audio is corrupted, i.e. the attention mechanism is not learning to correlate audio and video. In contrast, our Adapted-AVSR succeeds in learning monotonic AV alignments, crucial to exploit visual cues, even under noise.

We explored the minimum number of samples that are sufficient to make the system converge. In Fig. 6.3-(b), we show 4 Adapted-AVSR systems from speaker 31F where the number of sentences P^2 used for adaptation is reduced from 60 to 30, 20 and 10. The system converges to a similar point for all cases except 10 sentences, where there is no alignment and the behavior becomes similar to the ASR system shown in Fig. 6.3-(a). We attribute this robustness to the fact that sx are phonetically balanced sentences, and thus a limited number of sentences contains sufficient variety of mouth positions.

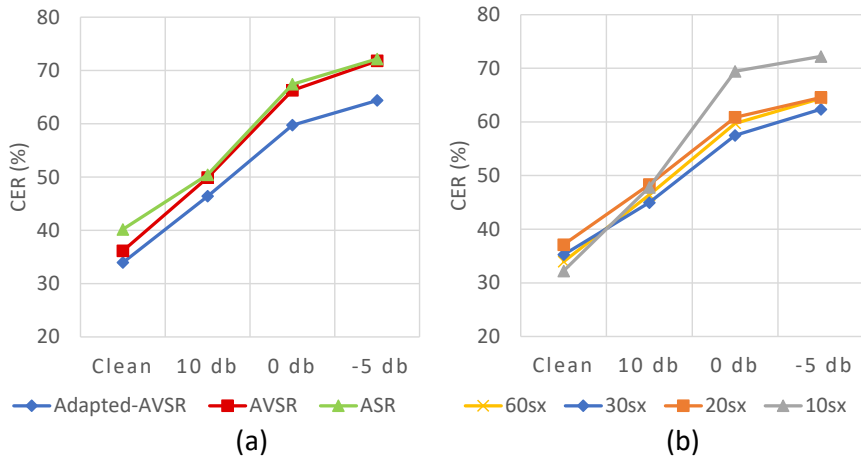


Figure 6.3: (a) CER on SI partition of TCD-TIMIT and (b) Adapted-AVSR with reduced P^2 .

6.4 Discussion and Conclusions

We investigate the adaptation of the visual front-end of an SI-AVSR system to an unseen and unlabeled speaker. We jointly tackle the speech recognition and the speaker adaptation problems. We use CoGANs to learn a joint distribution between corresponding lip-images that are used to minimize the inter-subject distances in the feature space. We test our system on TCD-TIMIT and achieve a 10% overall CER improvement with respect to the non-adapted audio-visual baseline. To the best of our knowledge, we are the first to propose speaker adaptation of the visual part of AVSR systems.

Compared to other speaker adaptation systems, targeting the audio domain, our work requires significantly less data and no labels. For example, Weninger et al. [235] adapt a Seq2seq audio-only model to a new speaker by minimizing the distance between the output distributions of the SI-ASR and the Adapted-ASR, achieving excellent results but requiring supervised learning and the availability of about 20 hours of data from the target speaker. In contrast, our system was able to adapt to a new speaker with as little as 20 phonetically balanced sentences of less than 10 seconds each.

Chapter 7

CONCLUSIONS

7.1 Research summary

In this thesis, we have focused on learning meaningful visual representations for continuous lip-reading. To this end, we have presented different data-driven mechanisms to handle the main challenges in lip-reading.

In our first work, **Chapter 3**, we have investigated the convenience of targeting directly phonemes or alternatively visemes. In particular, we proposed the automatic construction of a Spanish phoneme-to-viseme mapping based on visual similarities between phonemes to maximize word recognition. Our intuition supported the fact that every single frame or small window represents a mouth position that is related to one or more phonemes, i.e. we cannot directly classify phonemes. We have learned through our experiments that, even though going through visemes may seem like a loss of information, there is no perceivable difference, in visual terms, between some phonemes. Therefore, it is more useful to learn to distinguish what we can actually see/read than to confuse our models trying to differentiate not distinguishable visual events such as phonemes. This fact and the higher word accuracy obtained when using phoneme-to-viseme mappings, justified the usefulness of visemes instead of the direct use of phonemes for traditional ALR systems.

Nevertheless, the poor recognition rates of traditional ALR systems made us wonder if they were related to an inappropriate or incomplete design of ALR systems or directly to an intrinsic limitation in visual information that causes the impossibility of perfect decoding of the spoken message. In **Chapter 4**, we collected the VLRD database, the largest audio-visual dataset in Spanish, appropriately designed with the aim to estimate the recognition rates achievable by human observers and by an automatic system under optimal and comparable conditions. Overall, the dataset covers 10,200 words in total (1,374 unique) and

its total duration is around 180 minutes. The sentences contained a phonetically balanced distribution of the Spanish language and were accurately phonetically annotated.

Our experiments on human lip-reading coincided with prior reports indicating that people can *read* around a 30% of the information from the lips, and the rest is filled-in from the context [165, 50]. We also tested the performance of participants grouped by their hearing condition to compare their lip-reading abilities and verify if these were superior for hearing-impaired subjects, as suggested in some studies; and we found that although hearing-impaired participants outperformed normal-hearings on the lip-reading task, the differences were not statistically significant. Thus, we found very good lip-readers in both groups, supporting the fact that people really use lip-reading to a different extent depending on their hearing capability or the acoustic conditions.

Our experiments on automatic lip-reading reported half of the performance achieved by humans, and suggested that the gap between human and automatic lip-reading is more related to the interpretability of the context than to the ability to solve mouth appearance. This fact was highlighted when analyzing deeper the outputs of the system, i.e. the sequences returned by humans always made some sense, which was not generally true for our traditional ALR system as it did not include higher-level constraints. Accordingly, the key for decoding continuous lip-reading remains in the proper modeling of visual ambiguities through the incorporation of short and long-term contexts.

Nevertheless, the available datasets suitable for continuous lip-reading in most of the languages tend to be small-scale, which complicates the training of competitive models. Thus, in **Chapter 5**, we revealed that it is possible to train competitive end-to-end ALR systems with challenging small-scale datasets as long as the appropriate restrictions are made to the learning process, especially in terms of the visual front-end objective. To this end, we investigated the appropriate labels to train the visual front-end, and hypothesized that the visual front-end should be trained in a self-supervised setting, allowing it to target its own *visual units*, which we define as *a collection of visually similar images constrained by linguistics*. We show that these visual units can be generated in a fully automatic manner and are informative enough about the mouth and lip position to facilitate meaningful learning of visual features.

Our experiments have shown that those visual units foster proper learning of visual features, which otherwise are unreachable due to the limited amount of training data. Additionally, we also presented a data augmentation technique based on the synthesis of new video sequences from appropriately combining characters-like sub-sequences from existing videos, which help to deal with the exposure to a very reduced speech variability.

The proposed system obtained a 40% improvement in terms of characters with

respect to the baseline; and achieved competitive performance with respect to the state-of-the-art ALR systems trained in English, but using significantly fewer data and resources. Our results opened the door to future research in continuous lip-reading in multiple languages, where small-scale datasets are already available or can be recorded with significantly fewer efforts.

Finally, in **Chapter 6**, we explored the integration of audio-visual cues for continuous speech recognition with the aim to relax acoustic-dominance and increase the contribution of the visual cues, especially in noisy scenarios. Motivated by the fact that every person has unique mouth movements, which results in visual speech being highly speaker-dependent; and that speaker-dependent systems are impractical because they require large amounts of training data with annotations for each specific speaker, we proposed to rely on unsupervised speaker adaptation. In particular, we proposed to explore the visual domain adaptation of a speaker-independent AVSR system to an unknown and unlabeled speaker. Our assumption was that in an ideal speaker-independent system, the same speech events should be represented in a similar way, independently of the speaker, or what is the same, similar lip positions should be equivalently encoded.

Then, we adapted an AVSR system trained in a multispeaker source domain, i.e. using an audio-visual dataset of multiple speakers with annotations, to finally decode samples in a target domain (unknown speaker) using only a few video samples without annotations from both the source and target domains. Considering that we ignore when the same lip positions from different speakers are being uttered, we integrated CoGANs to generate corresponding lip-images to minimize the inter-subject distances in the feature space.

The proposed system was able to achieve a 15% overall CER improvement with respect to the audio-based system and 10% with respect to the non-adapted audio-visual baseline. Our experiments highlighted that the visual adaptation of a new speaker benefits the contribution of the visual domain in a speaker-independent AVSR systems, especially under the presence of acoustic noise. To the best of our knowledge, we were the first to propose unsupervised speaker adaptation of the visual part of AVSR systems.

7.2 Discussion and future work

There is a popular belief that speech is something that we hear, but there is overwhelming evidence that the brain treats speech as something that we hear and see. Following this statement, and considering that there are around 5 hundred million people around the world with hearing issues, and that these numbers are increasing every year (e.g. we are terribly exposed to noisy environments and the

use of headphones for a long time and/or at very high volumes), in this thesis, we explored automatic lip-reading systems with the aim to decode continuous speech from visual cues alone.

From our experiments, we have learned that visual speech is highly uncertain, at specific time instants or small windows, e.g. several phonemes produce the same lip movements. However, we found a tendency to handle these visual ambiguities by appropriately encoding spatial information and modeling temporal information at different levels of context. Thus, similarly to audio-based systems, where there is extensive research in the most convenient encoding of phonemes, we also investigated the encoding of the distinguishable visual units. In this way, we explored appropriate modeling of visual representations with a preference for those visual features that encode similar lip positions equivalently, independently of the speaker. We presented 2 contributions where visual representation that properly encode different lip-positions were key to decode visual speech: i) to handle limited data availability (Chapter 5), or ii) to increase the contribution of the visual domain in AVSR systems (Chapter 6). In both cases, the proper restrictions on the visual front-end objective helped in the generalization of the visual features, and also avoided a freewill learning, where prior knowledge was used to improve the training.

Furthermore, we also have learned that the availability of large and variate linguistic content is fundamental to advance in the field. Our experiments highlighted that although the visual features might be properly learned even when the model is exposed to a very limited context, it will not be able to generalize and decode new utterances. Thus, when targeting natural speech, we need to ensure the decoding of any feasible word, and this can be managed through the exposure of our models to sentences with many different contexts, i.e. the larger the provided linguistic content, the more chances we have to decode the target words. Accordingly, lip synthesis becomes necessary when the available datasets are limited to ensure proper modeling of temporal dependencies and decoding of new input sequences.

In conclusion, when training ALR systems we should consider a proper modeling of visual cues and a large exposition to different levels of context.

To end up, our results open the door to future research on i) transfer learning between different languages, i.e. deeper analysis in language dependencies; ii) balance the contribution of AVSR systems through spatio-temporal adaptation to new speakers; iii) alternative data augmentation through natural lip-reading synthesis; iv) transfer learning or context integration from other sources such as acoustic signals or text, where the amount of annotated data is much higher.

Bibliography

- [1] ACCAPS (2016). ACCAPS federació d'associacions catalanes de pares i persones sordes. <http://www.acapps.org/web/>. Accessed: 2016-08-16.
- [2] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2018a). Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.
- [3] Afouras, T., Chung, J. S., and Zisserman, A. (2018b). The conversation: Deep audio-visual speech enhancement. In *Proceedings of Interspeech*.
- [4] Afouras, T., Chung, J. S., and Zisserman, A. (2018c). Deep lip reading: A comparison of models and an online application. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 3514–3518.
- [5] Afouras, T., Chung, J. S., and Zisserman, A. (2018d). Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.
- [6] Afouras, T., Chung, J. S., and Zisserman, A. (2020). Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE.
- [7] Almajai, I., Cox, S., Harvey, R., and Lan, Y. (2016). Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 2722–2726.
- [8] Amer, R., Nassar, J., Bendahan, D., Greenspan, H., and Ben-Eliezer, N. (2019). Automatic segmentation of muscle tissue and inter-muscular fat in thigh and calf mri images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 219–227. Springer.

- [9] Anina, I., Zhou, Z., Zhao, G., and Pietikäinen, M. (2015). OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *Proc. International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 1–5.
- [10] Antonakos, E., Roussos, A., and Zafeiriou, S. (2015). A survey on mouth modeling and analysis for sign language recognition. In *Proc. International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 1–7.
- [11] Assael, Y. M., Shillingford, B., Whiteson, S., and de Freitas, N. (2017). Lipnet: Sentence-level lipreading. In *Proc. GPU Technology Conference*.
- [12] Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 3444–3451.
- [13] Bailly-Baillié, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., et al. (2003). The BANCA database and evaluation protocol. In *Proc. International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 625–638.
- [14] Bear, H. L., Cox, S. J., and Harvey, R. W. (2015). Speaker-independent machine lip-reading with speaker-dependent viseme classifiers. In *Proc. Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing*, pages 190–195.
- [15] Bear, H. L. and Harvey, R. (2016). Decoding visemes: improving machine lip-reading. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 2009–2013.
- [16] Bear, H. L. and Harvey, R. (2017). Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, 95:40–67.
- [17] Bear, H. L., Harvey, R. W., and Lan, Y. (2017). Finding phonemes: improving machine lip-reading. In *Proc. Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing*.
- [18] Bear, H. L., Harvey, R. W., Theobald, B.-J., and Lan, Y. (2014). Which phoneme-to-viseme maps best improve visual-only computer lip-reading? In *Proc. International Symposium on Visual Computing*, pages 230–239.
- [19] Benezeth, Y., Bachman, G., Le-Jan, G., Souviraà-Labastie, N., and Bimbot, F. (2011). *BL-Database: A French audiovisual database for speech driven lip animation systems*. PhD thesis, INRIA.

- [20] Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *Neural Networks*, 5(2):157–166.
- [21] Bernstein, L. E., Demorest, M. E., and Tucker, P. E. (1998). *What makes a good speechreader? First you have to find one*. Hove, United Kingdom: Psychology Press Ltd. Publishers.
- [22] Biswas, A., Sahu, P. K., and Chandra, M. (2015). Multiple camera in car audio-visual speech recognition using phonetic and visemic information. *Computers & Electrical Engineering*, 47:35–50.
- [23] Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., and Yvert, B. (2016). Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLoS Computational Biology*, 12(11):e1005119.
- [24] Boersma, P. et al. (2002). Praat, a system for doing phonetics by computer. *Glott international*, 5(9/10):341–345.
- [25] Bowden, R. (2010). LILiR language independent lip reading. <http://www.ee.surrey.ac.uk/Projects/LILiR/datasets.html>. Accessed: 2016-08-16.
- [26] Bowden, R., Cox, S., Harvey, R., Lan, Y., Ong, E.-J., Owen, G., and Theobald, B.-J. (2013). Recent developments in automated lip-reading. In *Optics and Photonics for Counterterrorism, Crime Fighting and Defence IX; and Optical Materials and Biomaterials in Security and Defence Systems Technology X*, volume 8901, page 89010J. International Society for Optics and Photonics.
- [27] Bozkurt, E., Erdem, C. E., Erzin, E., Erdem, T., and Ozkan, M. (2007). Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. In *Proc. International Conference on Signal Processing and Communications Applications*, pages 1–4.
- [28] Britto Mattos, A. and Borges Oliveira, D. A. (2018). Multi-view mouth renderization for assisting lip-reading. In *Proc. International Conference on the Web for All*.
- [29] Buchan, J. N., Paré, M., and Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, 2(1):1–13.
- [30] Burton, J., Frank, D., Saleh, M., Navab, N., and Bear, H. L. (2018). The speaker-independent lipreading play-off; a survey of lipreading machines. *Proc. ICIP*.

- [31] Capek, C. M., MacSweeney, M., Woll, B., Waters, D., McGuire, P. K., David, A. S., Brammer, M. J., and Campbell, R. (2008). Cortical circuits for silent speechreading in deaf and hearing people. *Neuropsychologia*, 46:1233–1241.
- [32] Cappelletta, L. and Harte, N. (2011). Viseme definitions comparison for visual-only speech recognition. In *Proc. European Conference on Signal Processing*, pages 2109–2113.
- [33] Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC*.
- [34] Chen, W., Tan, X., Xia, Y., Qin, T., Wang, Y., and Liu, T.-Y. (2020a). Duallip: A system for joint lip reading and generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1985–1993.
- [35] Chen, X., Du, J., and Zhang, H. (2020b). Lipreading with densenet and resbi-lstm. *Signal, Image and Video Processing*, pages 1–9.
- [36] Cheng, S., Ma, P., Tzimiropoulos, G., Petridis, S., Bulat, A., Shen, J., and Pantic, M. (2020). Towards pose-invariant lip-reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4357–4361. IEEE.
- [37] Chițu, A. and Rothkrantz, L. J. (2012). Automatic visual speech recognition. *Speech Enhancement, Modeling and Recognition—Algorithms and Applications*, pages 95–120.
- [38] Chitu, A. G., Driel, K., and Rothkrantz, L. J. (2010). Automatic lip reading in the Dutch language using active appearance models on high speed recordings. In *Proc. International Conference on Text, Speech and Dialogue*, pages 259–266.
- [39] Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., et al. (2018). State-of-the-art speech recognition with sequence-to-sequence models. In *Proc. ICASSP*, pages 4774–4778.
- [40] Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *Proc. Interspeech 2018*, pages 1086–1090.
- [41] Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 3444–3453.

- [42] Chung, J. S. and Zisserman, A. (2016a). Lip reading in the wild. In *Proc. Asian Conference on Computer Vision*, pages 87–103.
- [43] Chung, J. S. and Zisserman, A. (2016b). Out of time: automated lip sync in the wild. In *Proc. Asian Conference on Computer Vision*, pages 251–263.
- [44] Chung, J. S. and Zisserman, A. (2017). Lip reading in profile. In *Proc. British Machine Vision Conference*.
- [45] Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Open Science*, 1:140216.
- [46] Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(5):2421–2424.
- [47] Cox, S. J., Harvey, R., Lan, Y., Newman, J. L., and Theobald, B.-J. (2008). The challenge of multispeaker lip-reading. In *Proc. International Conference on Auditory-Visual Speech Processing*, pages 179–184.
- [48] Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., and Szykalski, M. (2017). An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems*, pages 1–26.
- [49] Davis, K. H., Biddulph, R., and Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642.
- [50] Duchnowski, P., Lum, D. S., Krause, J. C., Sexton, M. G., Bratakos, M. S., and Braida, L. D. (2000). Development of speechreading supplements based on automatic speech recognition. *Biomedical Engineering*, 47(4):487–496.
- [51] Dupont, S. and Luetin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151.
- [52] Ellis, T., MacSweeney, M., Dodd, B., and Campbell, R. (2001). Tas: A new test of adult speechreading-deaf people really can be better speechreaders. In *AVSP*.
- [53] Ephrat, A., Halperin, T., and Peleg, S. (2017). Improved speech reconstruction from silent video. In *Proc. International Workshop on Computer Vision for Audio-Visual Media*.

- [54] Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, 40(4):481–492.
- [55] Estellers, V., Gurban, M., and Thiran, J.-P. (2012). On dynamic stream weighting for audio-visual speech recognition. *IEEE-ACM Transactions on Audio, Speech, and Language Processing*, 20(4):1145–1157.
- [56] Estellers, V. and Thiran, J.-P. (2011). Multipose audio-visual speech recognition. In *Proc. European Conference on Signal Processing*, pages 1065–1069.
- [57] Estellers, V. and Thiran, J.-P. (2012). Multi-pose lipreading and audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2012(1):51.
- [58] Estival, D., Cassidy, S., Cox, F., and Burnham, D. (2014). AusTalk: an audio-visual corpus of Australian English. In *Proc. International Conference on Language Resources and Evaluation*.
- [59] Eveno, N., Caplier, A., and Coulon, P.-Y. (2004). Accurate and quasi-automatic lip tracking. *Circuits and Systems for Video Technology*, 14(5):706–715.
- [60] Ezzat, T. and Poggio, T. (1998). Miketalk: A talking facial display based on morphing visemes. In *Proc. Conference on Computer Animation*, pages 96–102.
- [61] Fernandez-Lopez, A., Martinez, O., and Sukno, F. M. (2017). Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database. In *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 208–215.
- [62] Fernandez-Lopez, A. and Sukno, F. (2018). Survey on automatic lip-reading in the era of deep learning. *Image Vision Comput.*, 78:53–72.
- [63] Fernandez-Lopez, A. and Sukno, F. M. (2017a). Automatic viseme vocabulary construction to enhance continuous lip-reading. *Proc. International Conference on Computer Vision Theory and Applications*, 5:52–63.
- [64] Fernandez-Lopez, A. and Sukno, F. M. (2017b). Automatic viseme vocabulary construction to enhance continuous lip-reading. *VISAPP*.
- [65] Fernandez-Lopez, A. and Sukno, F. M. (2019). Lip-reading with limited-data network. In *27th European Signal Processing Conference, EUSIPCO 2019, A Coruña, Spain, September 2-6, 2019*.

- [66] Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research*, 11(4):796–804.
- [67] Fox, N. A., O’Mullane, B. A., and Reilly, R. B. (2005). VALID: A new practical audio-visual database, and comparative results. In *Proc. International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 777–786.
- [68] Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., and Fralish, J. S. (1995). Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, 6(1):99–106.
- [69] Frénay, B. and Verleysen, M. (2014). Classification in the presence of label noise: a survey. *Neural Networks and Learning Systems*, 25(5):845–869.
- [70] Fu, Y., Zhou, X., Liu, M., Hasegawa-Johnson, M., and Huang, T. S. (2007). Lipreading by locality discriminant graph. In *Proc. International Conference on Image Processing*, volume 3, pages 325–328.
- [71] Fung, H. L. and Mak, B. (2018). End-to-end low-resource lip-reading with maxout CNN and LSTM. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.
- [72] Gabbay, A., Ephrat, A., Halperin, T., and Peleg, S. (2017). Seeing through noise: Speaker separation and enhancement using visually-derived speech. In *Proc. International Workshop on Computer Vision for Audio-Visual Media*.
- [73] Gabbay, A., Shamir, A., and Peleg, S. (2018). Visual speech enhancement. In *Proceedings of Interspeech*.
- [74] Georgakis, C., Petridis, S., and Pantic, M. (2014a). Discriminating native from non-native speech using fusion of visual cues. In *Proc. International Conference on Multimedia*, pages 1177–1180.
- [75] Georgakis, C., Petridis, S., and Pantic, M. (2014b). Visual-only discrimination between native and non-native speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 4828–4832. IEEE.
- [76] Georgakis, C., Petridis, S., and Pantic, M. (2016). Discrimination between native and non-native speech using visual features only. *IEEE Transactions on Cybernetics*, 46(12):2758–2771.

- [77] Gers, F. A., Schmidhuber, J. A., and Cummins, F. A. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471.
- [78] Goecke, R. and Millar, J. B. (2004). The audio-video australian english speech data corpus AVOZES. In *Proc. International Conference on Spoken Language Processing*, pages 2525–2528.
- [79] Goldschen, A. J., Garcia, O. N., and Petajan, E. (1994). Continuous optical automatic speech recognition by lipreading. In *Proc. Conference on Signals, Systems and Computers*, volume 1, pages 572–577.
- [80] Gowdy, J. N., Subramanya, A., Bartels, C., and Bilmes, J. (2004). DBN based multi-stream models for audio-visual speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–993.
- [81] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. International Conference on Machine Learning*, pages 369–376.
- [82] Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *Proc. International Conference on Machine Learning*, volume 14, pages 1764–1772.
- [83] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610.
- [84] Gurban, M. and Thiran, J.-P. (2009). Information theoretic feature extraction for audio-visual speech recognition. *Signal Processing*, 57(12):4765–4776.
- [85] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. In *Proc. International Conference on Machine Learning*.
- [86] Harte, N. and Gillen, E. (2015). TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615.
- [87] Hazen, T. J., Saenko, K., La, C.-H., and Glass, J. R. (2004). A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *Proc. International Conference on Multimodal Interfaces*, pages 235–242.

- [88] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- [89] Hilder, S., Harvey, R., and Theobald, B.-J. (2009). Comparison of human and machine-based lip-reading. In *Proc. International Conference on Auditory-Visual Speech Processing*, pages 86–89.
- [90] Hoffer, E. and Ailon, N. (2015). Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer.
- [91] Hong, X., Yao, H., Wan, Y., and Chen, R. (2006). A PCA based visual DCT feature extraction method for lip-reading. In *Proc. International Conference on Intelligent Information Hiding and Multimedia*, pages 321–326.
- [92] Howell, D., Cox, S., and Theobald, B. (2016). Visual units and confusion modelling for automatic lip-reading. *Image and Vision Computing*, 51:1–12.
- [93] Howell, D. L. (2015). *Confusion modelling for lip-reading*. PhD thesis, University of East Anglia.
- [94] Hu, D., Li, X., et al. (2016). Temporal multimodal learning in audiovisual speech recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 3574–3582.
- [95] Huang, J. and Kingsbury, B. (2013). Audio-visual deep learning for noise robust speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 7596–7599.
- [96] Huang, J., Potamianos, G., Connell, J., and Neti, C. (2004). Audio-visual speech recognition using an infrared headset. *Speech Communication*, 44(1):83–96.
- [97] Igras, M., Ziółko, B., and Jadczyk, T. (2012). Audiovisual database of Polish speech recordings. *Studia Informatica*, 33(2B):163–172.
- [98] Jaumard-Hakoun, A., Xu, K., Leboullenger, C., Roussel-Ragot, P., and Denby, B. (2016). An articulatory-based singing voice synthesis using tongue and lips imaging. In *Proceedings of Interspeech*, pages 1467–1471.
- [99] Jeffers, J. and Barley, M. (1971). *Speechreading (lipreading)*. Thomas.

- [100] Kailath, T. (1967). The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60.
- [101] Kandala, P. A., Thanda, A., Margam, D. K., Aralikatti, R. C., Sharma, T., Roy, S., and Venkatesan, S. M. (2019). Speaker adaptation for lip-reading using visual identity vectors. In *INTERSPEECH*, pages 2758–2762.
- [102] Kannan, A., Wu, Y., Nguyen, P., Sainath, T. N., Chen, Z., and Prabhavalkar, R. (2018). An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5828. IEEE.
- [103] Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A. (2011). Comparing boosting and bagging techniques with noisy and imbalanced data. *Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(3):552–568.
- [104] Kolossa, D., Zeiler, S., Vorwerk, A., and Orglmeister, R. (2009). Audiovisual speech recognition with missing or unreliable data. In *Proc. International Conference on Auditory-Visual Speech Processing*, pages 117–122.
- [105] Koumparoulis, A. and Potamianos, G. (2018). Deep view2view mapping for view-invariant lipreading. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 588–594. IEEE.
- [106] Koumparoulis, A. and Potamianos, G. (2019). Mobilipnet: Resource-efficient deep learning based lipreading. In *INTERSPEECH*, pages 2763–2767.
- [107] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proc. Conference on Advances in Neural Information Processing Systems*, pages 1097–1105.
- [108] Kumar, K., Chen, T., and Stern, R. M. (2007). Profile view lip reading. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 429–432.
- [109] Kuwabara, H., Takeda, K., Sagisaka, Y., Katagiri, S., Morikawa, S., and Watanabe, T. (1989). Construction of a large-scale Japanese speech database and its management system. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 560–563.
- [110] Kyle, F. E., Campbell, R., Mohammed, T., Coleman, M., and MacSweeney, M. (2013). Speechreading development in deaf and hearing children:

- introducing the test of child speechreading. *J Speech Lang Hear Res*, 56:416–426.
- [111] Kyle, F. E. and Harris, M. (2006). Concurrent correlates and predictors of reading and spelling achievement in deaf and hearing school children. *J Deaf Stud Deaf Educ*, 11:273–288.
- [112] Lan, Y., Harvey, R., Theobald, B., Ong, E.-J., and Bowden, R. (2009). Comparing visual features for lipreading. In *Proc. International Conference on Auditory-Visual Speech Processing*, pages 102–106.
- [113] Lan, Y., Harvey, R., and Theobald, B.-J. (2012a). Insights into machine lip reading. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 4825–4828.
- [114] Lan, Y., Theobald, B.-J., and Harvey, R. (2012b). View independent computer lip-reading. In *Proc. International Conference on Multimedia and Expo*, pages 432–437.
- [115] Lan, Y., Theobald, B.-J., Harvey, R., Ong, E.-J., and Bowden, R. (2010). Improving visual features for lip-reading. In *Proc. International Conference on Auditory-Visual Speech Processing*.
- [116] Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., and Huang, T. S. (2004). AVICAR: audio-visual speech corpus in a car environment. In *Proceedings of Interspeech*.
- [117] Lee, D., Lee, J., and Kim, K.-E. (2016). Multi-view automatic lip-reading using neural network. In *Proc. Asian Conference on Computer Vision*, pages 290–302.
- [118] Lesani, F. S., Ghazvini, F. F., and Dianat, R. (2015). Mobile phone security using automatic lip reading. In *Proc. International Conference on E-Commerce in Developing Countries: With focus on e-Business*, pages 1–5.
- [119] Leung, K.-Y., Mak, M.-W., and Kung, S.-Y. (2004). Articulatory feature-based conditional pronunciation modeling for speaker verification. In *Proc. ICSLP*.
- [120] Lin, M., Chen, Q., and Yan, S. (2014). Network in network. In *Proc. International Conference on Learning Representations*.
- [121] Liu, M.-Y. and Tuzel, O. (2016). Coupled generative adversarial networks. In *Proc. NIPS*, pages 469–477.

- [122] Llisterri, J. and Mariño, J. B. (1993). Spanish adaptation of sampa and automatic phonetic transcription. *Reporte técnico del Espirit Project*, 6819.
- [123] Lott, B. E. and Levy, J. (1960). The influence of certain communicator characteristics on lip reading efficiency. *J. Appl. Soc. Psychol.*, 51(2):419–425.
- [124] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.
- [125] Lu, L., Yu, J., Chen, Y., Liu, H., Zhu, Y., Liu, Y., and Li, M. (2018). Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1466–1474. IEEE.
- [126] Lucey, P. and Potamianos, G. (2006). Lipreading using profile versus frontal views. In *Proc. International Workshop on Multimedia Signal Processing*, pages 24–28.
- [127] Lucey, P. J., Potamianos, G., and Sridharan, S. (2007). A unified approach to multi-pose audio-visual ASR. In *Proceedings of Interspeech*, pages 650–653.
- [128] Lucey, P. J., Potamianos, G., and Sridharan, S. (2008a). Patch-based analysis of visual speech from multiple views. In *Proc. International Conference on Auditory-Visual Speech Processing*.
- [129] Lucey, P. J., Sridharan, S., and Dean, D. B. (2008b). Continuous pose-invariant lipreading. In *Proceedings of Interspeech*, pages 2679–2682.
- [130] Luettin, J., Thacker, N. A., and Beet, S. W. (1996). Visual speech recognition using active shape models and hidden markov models. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 817–820.
- [131] Luo, M., Yang, S., Shan, S., and Chen, X. (2020). Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 69–76.
- [132] Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

- [133] Lyxell, B. and Holmberg, I. (2000). Visual speechreading and cognitive performance in hearing-impaired and normal hearing children (11-14 years). *Br J Educ Psychol*, 70:505–518.
- [134] Makino, T., Liao, H., Assael, Y., Shillingford, B., Garcia, B., Braga, O., and Siohan, O. (2019). Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 905–912. IEEE.
- [135] Marcheret, E., Libal, V., and Potamianos, G. (2007). Dynamic stream weight modeling for audio-visual speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 945–948.
- [136] Martinez, B., Ma, P., Petridis, S., and Pantic, M. (2020). Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE.
- [137] Mase, K. and Pentland, A. (1991). Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22(6):67–76.
- [138] Mathias, M., Benenson, R., Pedersoli, M., and Van Gool, L. (2014). Face detection without bells and whistles. In *Proc. European Conference on Computer Vision*, pages 720–735.
- [139] Mathulaprangsan, S., Wang, C.-Y., Kusum, A. Z., Tai, T.-C., and Wang, J.-C. (2015). A survey of visual lip reading and lip-password verification. In *Proc. International Conference on Orange Technologies*, pages 22–25.
- [140] Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., and Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213.
- [141] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: trainable text-speech alignment using kald. In *Proc. of interspeech*, pages 498–502.
- [142] McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matejka, P., Cernocký, J., Poh, N., Kittler, J., Larcher, A., Levy, C., et al. (2012). Bi-modal person recognition on a mobile phone: using mobile phone data. In *Proc. International Workshop on Multimedia and Expo*, pages 635–640.

- [143] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.
- [144] Messer, K., Matas, J., Kittler, J., Luetin, J., and Maitre, G. (1999). XM2VTSDB: The extended M2VTS database. In *Proc. International Conference on Audio and Video-based Biometric Person Authentication*, volume 964, pages 965–966.
- [145] Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., and Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clin Linguist Phon*, 20:621–630.
- [146] Moll, K. L. and Daniloff, R. G. (1971). Investigation of the timing of velar movements during speech. *Journal of the Acoustical Society of America*, 50(2B):678–684.
- [147] Moon, S., Kim, S., and Wang, H. (2015). Multimodal transfer deep learning with applications in audio-visual recognition. *MMML Workshop at Neural Information Processing Systems*.
- [148] Mroueh, Y., Marcheret, E., and Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 2130–2134.
- [149] Navarathna, R., Kleinschmidt, T., Dean, D. B., Sridharan, S., and Lucey, P. J. (2011). Can audio-visual speech recognition outperform acoustically enhanced speech recognition in automotive environment? In *Proceedings of Interspeech*, pages 2241–2244.
- [150] Nefian, A. V., Liang, L., Pi, X., Liu, X., and Murphy, K. (2002a). Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002(11):1–15.
- [151] Nefian, A. V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., and Murphy, K. (2002b). A coupled HMM for audio-visual speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 2013–2016.
- [152] Neti, C., Potamianos, G., Luetin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., and Mashari, A. (2000). Audio visual speech recognition. Technical report, IDIAP.
- [153] Nettleton, D. F., Orriols-Puig, A., and Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33(4):275–306.

- [154] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proc. International Conference on Machine Learning*, pages 689–696.
- [155] Niklaus, S., Mai, L., and Liu, F. (2017). Video frame interpolation via adaptive separable convolution. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 261–270.
- [156] Ninomiya, H., Kitaoka, N., Tamura, S., Iribe, Y., and Takeda, K. (2015). Integration of deep bottleneck features for audio-visual speech recognition. In *Proceedings of Interspeech*, pages 563–567.
- [157] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., and Ogata, T. (2014). Lipreading using convolutional neural network. In *Proceedings of Interspeech*, pages 1149–1153.
- [158] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., and Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42(4):722–737.
- [159] Ong, E.-J. and Bowden, R. (2011a). Learning sequential patterns for lipreading. In *Proc. British Machine Vision Conference*.
- [160] Ong, E.-J. and Bowden, R. (2011b). Learning temporal signatures for lip reading. In *Proc. International Conference on Computer Vision Workshops*, pages 958–965.
- [161] Ordóñez, F. J. and Roggen, D. (2016). Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115.
- [162] Orozco, J., Martinez, B., and Pantic, M. (2015). Empirical analysis of cascade deformable models for multi-view face detection. *Image and Vision Computing*, 42:47–61.
- [163] Orozco, J., Rudovic, O., González, J., and Pantic, M. (2013). Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. *Image and Vision Computing*, 31(4):322–340.
- [164] Ortega, A., Sukno, F., Lleida, E., Frangi, A. F., Miguel, A., Buera, L., and Zacur, E. (2004). AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In *Proc. International Conference on Language Resources and Evaluation*, pages 763–767.

- [165] Ortiz, I. d. I. R. R. (2008). Lipreading in the prelingually deaf: what makes a skilled speechreader? *The Spanish Journal of Psychology*, 11(02):488–502.
- [166] Pachoud, S., Gong, S., and Cavallaro, A. (2008). Macro-cuboid based probabilistic matching for lip-reading digits. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 1–8.
- [167] Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P. (2008). Adaptive multimodal fusion by uncertainty compensation with application to audio-visual speech recognition. In *Proc. International Conference on Multimodal Processing and Interaction*, pages 1–15.
- [168] Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P. (2009). Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE-ACM Transactions on Audio, Speech, and Language Processing*, 17(3):423–435.
- [169] Pass, A., Zhang, J., and Stewart, D. (2010). An investigation into features for multi-view lipreading. In *Proc. International Conference on Image Processing*, pages 2417–2420.
- [170] Patterson, E. K., Gurbuz, S., Tufekci, Z., and Gowdy, J. N. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 2017–2020.
- [171] Pei, Y., Kim, T.-K., and Zha, H. (2013). Unsupervised random forest manifold alignment for lipreading. In *Proc. IEEE International Conference on Computer Vision*, pages 129–136.
- [172] Petridis, S., Li, Z., and Pantic, M. (2017a). End-to-end visual speech recognition with LSTMs. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 2592–2596.
- [173] Petridis, S. and Pantic, M. (2016). Deep complementary bottleneck features for visual speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 2304–2308.
- [174] Petridis, S., Shen, J., Cetin, D., and Pantic, M. (2018a). Visual-only recognition of normal, whispered and silent speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.
- [175] Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., and Pantic, M. (2018b). End-to-end audiovisual speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.

- [176] Petridis, S., Wang, Y., Li, Z., and Pantic, M. (2017b). End-to-end audiovisual fusion with LSTMs. In *Proc. International Conference on Auditory-Visual Speech Processing*.
- [177] Petridis, S., Wang, Y., Li, Z., and Pantic, M. (2017c). End-to-end multi-view lipreading. In *Proc. British Machine Vision Conference*.
- [178] Petridis, S., Wang, Y., Ma, P., Li, Z., and Pantic, M. (2020). End-to-end visual speech recognition for small-scale datasets. *Pattern Recognition Letters*, 131:421–427.
- [179] Petrovska-Delacrétaz, D., Lelandais, S., Colineau, J., Chen, L., Dorizzi, B., Ardabilian, M., Krichen, E., Mellakh, M.-A., Chaari, A., Guerfi, S., et al. (2008). The IV 2 multimodal biometric database (including iris, 2D, 3D, stereoscopic, and talking face data), and the IV 2-2007 evaluation campaign. In *Proc. International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7.
- [180] Petrushin, V. A. (2000). Hidden markov models: Fundamentals and applications. In *OSEE*.
- [181] Polyakova, T. V. (2015). Grapheme-to-phoneme conversion in the era of globalization. *Thesis*.
- [182] Potamianos, G. and Neti, C. (2001). Automatic speechreading of impaired speech. In *Proc. International Conference on Auditory-Visual Speech Processing*.
- [183] Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326.
- [184] Potamianos, G., Neti, C., Luetin, J., and Matthews, I. (2004). Audio-visual automatic speech recognition: An overview. *Issues in Visual and Audio-Visual Speech Processing*, 22:23–61.
- [185] Qu, L., Weber, C., and Wermter, S. (2019). Lipsound: Neural mel-spectrogram reconstruction for lip reading. In *INTERSPEECH*, pages 2768–2772.
- [186] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

- [187] Rahmani, M. H. and Almasganj, F. (2017). Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features. In *Proc. International Conference on Pattern Recognition and Image Analysis*, pages 195–199.
- [188] Rekik, A., Ben-Hamadou, A., and Mahdi, W. (2014). A new visual speech recognition approach for RGB-D cameras. In *Proc. International Conference on Image Analysis and Recognition*, pages 21–28.
- [189] Rekik, A., Ben-Hamadou, A., and Mahdi, W. (2016). An adaptive approach for lip-reading using image and depth data. *Multimedia Tools and Applications*, 75(14):8609–8636.
- [190] Rodríguez-Ortiz, I. R., Saldaña, D., and Moreno-Perez, F. J. (2015). How speechreading contributes to reading in a transparent orthography: the case of spanish deaf people. *J Res Read*.
- [191] Ronquest, R. E., Levi, S. V., and Pisoni, D. B. (2010). Language identification from visual-only speech signals. *Attention, Perception, & Psychophysics*, 72(6):1601–1613.
- [192] Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, 17(6):405–409.
- [193] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18.
- [194] Sahu, V. and Sharma, M. (2013). Result based analysis of various lip tracking systems. In *Proc. International Conference on Green High Performance Computing*, pages 1–7.
- [195] Saitoh, T. and Konishi, R. (2010). Profile lip reading for vowel and word recognition. In *Proc. International Conference on Pattern Recognition*, pages 1356–1359.
- [196] Saitoh, T., Zhou, Z., Zhao, G., and Pietikäinen, M. (2016). Concatenated frame image based CNN for visual speech recognition. In *Proc. Asian Conference on Computer Vision*, pages 277–289.
- [197] Salik, K. M., Aggarwal, S., Kumar, Y., Shah, R. R., Jain, R., and Zimmermann, R. (2019). Lipper: Speaker independent speech synthesis using multi-view lipreading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 10023–10024.

- [198] Sanderson, C. (2002). The VidTIMIT database. Technical report, IDIAP.
- [199] Sengupta, S., Bhattacharya, A., Desai, P., and Gupta, A. (2012). Automated lip reading technique for password authentication. *International Journal of Applied Information Systems*, pages 2249–0868.
- [200] Seymour, R., Stewart, D., and Ming, J. (2008). Comparison of image transform-based features for visual speech recognition in clean and corrupted videos. *Journal on Signal Image and Video Processing*, pages 14–22.
- [201] Shao, X. and Barker, J. (2008). Stream weight estimation for multistream audio–visual speech recognition in a multispeaker environment. *Speech Communication*, 50(4):337–353.
- [202] Shillingford, B., Assael, Y., Hoffman, M. W., Paine, T., Hughes, C., Prabhu, U., Liao, H., Sak, H., Rao, K., Bennett, L., Mulville, M., Coppin, B., Laurie, B., Senior, A., and de Freitas, N. (2019a). Large-scale visual speech recognition.
- [203] Shillingford, B., Assael, Y. M., Hoffman, M. W., Paine, T., Hughes, C., Prabhu, U., Liao, H., Sak, H., Rao, K., Bennett, L., Mulville, M., Denil, M., Coppin, B., Laurie, B., Senior, A. W., and de Freitas, N. (2019b). Large-scale visual speech recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 4135–4139. ISCA.
- [204] Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385.
- [205] Stafylakis, T. and Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. In *Proceedings of Interspeech*, pages 3652–3656.
- [206] Sterpu, G. and Harte, N. (2017). Towards lipreading sentences using active appearance models. In *Proc. International Conference on Auditory-Visual Speech Processing*.
- [207] Sterpu, G., Saam, C., and Harte, N. (2018). Attention-based audio-visual fusion for robust automatic speech recognition. In *Proc. ICMI*, pages 111–115.
- [208] Sterpu, G., Saam, C., and Harte, N. (2020). How to teach dnns to pay attention to the visual modality in speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1052–1064.

- [209] Stewart, D., Seymour, R., Pass, A., and Ming, J. (2014). Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE Transactions on Cybernetics*, 44(2):175–184.
- [210] Sui, C., Bennamoun, M., and Togneri, R. (2015). Listening with your eyes: Towards a practical visual speech recognition system using deep boltzmann machines. In *Proc. International Conference on Computer Vision*, pages 154–162.
- [211] Sui, C., Togneri, R., and Bennamoun, M. (2017). A cascade gray-stereo visual feature extraction method for visual and audio-visual speech recognition. *Speech Communication*, 90:26–38.
- [212] Sukno, F. M., Ordas, S., Butakoff, C., Cruz, S., and Frangi, A. F. (2007). Active shape models with invariant optimal features: application to facial analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1105–1117.
- [213] Sumbly, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2):212–215.
- [214] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 1–9.
- [215] Takashima, Y., Aihara, R., Takiguchi, T., Arika, Y., Mitani, N., Omori, K., and Nakazono, K. (2016). Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss. In *Proceedings of Interspeech*, pages 277–281.
- [216] Tamura, S., Miyajima, C., Kitaoka, N., Yamada, T., Tsuge, S., Takiguchi, T., Yamamoto, K., Nishiura, T., Nakayama, M., Denda, Y., et al. (2010). CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition. In *Proc. International Conference on Auditory-Visual Speech Processing*.
- [217] Tan, J., Wang, X., Nguyen, C.-T., and Shi, Y. (2018). Silentkey: A new authentication framework through ultrasonic-based lip reading. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–18.

- [218] Thangthai, K., Bear, H. L., and Harvey, R. (2017). Comparing phonemes and visemes with DNN-based lipreading. In *Proc. British Machine Vision Conference*.
- [219] Thangthai, K. and Harvey, R. (2017). Improving computer lipreading via DNN sequence discriminative training techniques. *Proceedings of Interspeech*, pages 3657–3661.
- [220] Thangthai, K., Harvey, R., Cox, S., and Theobald, B.-J. (2015). Improving lip-reading performance for robust audiovisual speech recognition using DNNs. In *Proc. International Conference on Auditory-Visual Speech Processing*, pages 127–131.
- [221] Trojanová, J., Hruží, M., Campr, P., and Železný, M. (2008). Design and recording of Czech audio-visual database with impaired conditions for continuous speech recognition. In *Proc. International Conference on Language Resources and Evaluation*.
- [222] Twaddell, W. F. (1935). On defining the phoneme. *Language*, 11(1):5–62.
- [223] Tzimiropoulos, G. and Pantic, M. (2017). Fast algorithms for fitting active appearance models to unconstrained images. *International Journal of Computer Vision*, 122(1):17–33.
- [224] Verbaeten, S. and Van Assche, A. (2003). Ensemble methods for noise elimination in classification problems. In *Proc. International Workshop on Multiple Classifier Systems*, pages 317–325.
- [225] Verkhodanova, V., Ronzhin, A., Kipyatkova, I., Ivanko, D., Karpov, A., and Železný, M. (2016). HAVRUS corpus: high-speed recordings of audio-visual Russian speech. In *Proc. International Conference on Speech and Computer*, pages 338–345.
- [226] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.
- [227] Vorwerk, A., Wang, X., Kolossa, D., Zeiler, S., and Orglmeister, R. (2010). WAPUSK20 - A database for robust audiovisual speech recognition. In *Proc. International Conference on Language Resources and Evaluation*.
- [228] Wand, M., Koutník, J., and Schmidhuber, J. (2016). Lipreading with long short-term memory. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 6115–6119.

- [229] Wand, M. and Schmidhuber, J. (2017). Improving speaker-independent lipreading with domain-adversarial training. In *Proceedings of Interspeech*, pages 3662–3666.
- [230] Wand, M., Vu, N. T., and Schmidhuber, J. (2018). Investigations on end-to-end audiovisual fusion. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.
- [231] Wang, C. (2019). Multi-grained spatio-temporal modeling for lip-reading. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 276. BMVA Press.
- [232] Wang, S.-L., Liew, A. W.-C., Lau, W. H., and Leung, S. H. (2008). An automatic lipreading system for spoken digits with limited training data. *Circuits and Systems for Video Technology*, 18(12):1760–1765.
- [233] Wells, J. C. et al. (1997). Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4.
- [234] Weng, X. and Kitani, K. (2019). Learning spatio-temporal features with two-stream deep 3d cnns for lipreading. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 269. BMVA Press.
- [235] Weninger, F., Andrés-Ferrer, J., Li, X., and Zhan, P. (2019). Listen, Attend, Spell and Adapt: Speaker Adapted Sequence-to-Sequence ASR. *Proc. Interspeech*.
- [236] Williams, J. J., Rutledge, J. C., Katsaggelos, A. K., and Garstecki, D. C. (1998). Frame rate and viseme analysis for multimedia applications to assist speechreading. *Journal of Signal Processing Systems*, 20(1-2):7–23.
- [237] Wong, Y. W., Ch’ng, S. I., Seng, K. P., Ang, L.-M., Chin, S. W., Chew, W. J., and Lim, K. H. (2011). A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities. *Pattern Recognition Letters*, 32(13):1503–1510.
- [238] Wu, P., Liu, H., Li, X., Fan, T., and Zhang, X. (2016). A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. *IEEE Transactions on Multimedia*, 18(3):326–338.
- [239] Wu, Y., Hassner, T., Kim, K., Medioni, G., and Natarajan, P. (2017). Facial landmark detection with tweaked convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

- [240] Xiao, J., Yang, S., Zhang, Y., Shan, S., and Chen, X. (2020). Deformation flow based two-stream network for lip reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 836–842.
- [241] Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 532–539.
- [242] Xu, K., Li, D., Cassimatis, N., and Wang, X. (2018). Lcanet: End-to-end lipreading with cascaded attention-ctc. In *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 548–555.
- [243] Yakel, D. A., Rosenblum, L. D., and Fortier, M. A. (2000). Effects of talker variability on speechreading. *Atten Percept Psychophys*, 62(7):1405–1412.
- [244] Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., Long, K., Shan, S., and Chen, X. (2019). Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE.
- [245] Yao, X. L. H. and Wang, X. H. Q. (2008). HIT-AVDB-II: A new multi-view and extreme feature cases contained audio-visual database for biometrics. In *Proc. Conference on Information Sciences*.
- [246] Yau, W. C., Kumar, D. K., and Weghorn, H. (2007). Visual speech recognition using motion features and hidden markov models. In *Proc. International Conference on Computer Analysis of Images and Patterns*, pages 832–839.
- [247] Yu, X., Huang, J., Zhang, S., and Metaxas, D. N. (2016). Face landmark fitting via optimized part mixtures and cascaded deformable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2212–2226.
- [248] Zafeiriou, S., Zhang, C., and Zhang, Z. (2015). A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24.
- [249] Zhang, X., Cheng, F., and Wang, S. (2019). Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 713–722.

- [250] Zhao, G., Barnard, M., and Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265.
- [251] Zhao, X., Yang, S., Shan, S., and Chen, X. (2020a). Mutual information maximization for effective lip reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 843–850.
- [252] Zhao, Y., Xu, R., and Song, M. (2019). A cascade sequence-to-sequence model for chinese mandarin lip reading. In *Proceedings of the ACM Multimedia Asia*, pages 1–6.
- [253] Zhao, Y., Xu, R., Wang, X., Hou, P., Tang, H., and Song, M. (2020b). Hearing lips: Improving lip reading by distilling speech recognizers. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6917–6924. AAAI Press.
- [254] Zhou, Z., Hong, X., Zhao, G., and Pietikäinen, M. (2014a). A compact representation of visual speech data using latent variables. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1).
- [255] Zhou, Z., Zhao, G., Hong, X., and Pietikäinen, M. (2014b). A review of recent advances in visual speech decoding. *Image and Vision Computing*, 32(9):590–605.
- [256] Zhou, Z., Zhao, G., and Pietikainen, M. (2010). Lipreading: a graph embedding approach. In *Proc. International Conference on Pattern Recognition*, pages 523–526.
- [257] Zhou, Z., Zhao, G., and Pietikäinen, M. (2011). Towards a practical lipreading system. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 137–144.
- [258] Zimmermann, M., Ghazi, M. M., Ekenel, H. K., and Thiran, J.-P. (2016). Visual speech recognition using PCA networks and LSTMs in a tandem GMM-HMM system. In *Proc. Asian Conference on Computer Vision*, pages 264–276.