IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Analysis of Visual Features for Continuous Lipreading in Spanish

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

Pattern Recognition and Human Language Technologies Research Center,
Universitat Politècnica de València, Camino de Vera, s/n, 46022, València, Spain

dagigo1@dsic.upv.es, cmartine@dsic.upv.es

Abstract

During a conversation, our brain is responsible for combining information obtained from multiple senses in order to improve our ability to understand the message we are perceiving. Different studies have shown the importance of presenting visual information in these situations. Nevertheless, lipreading is a complex task whose objective is to interpret speech when audio is not available. By dispensing with a sense as crucial as hearing, it will be necessary to be aware of the challenge that this lack presents. In this paper, we propose an analysis of different speech visual features with the intention of identifying which of them is the best approach to capture the nature of lip movements for natural Spanish and, in this way, dealing with the automatic visual speech recognition task. In order to estimate our system, we present an audiovisual corpus compiled from a subset of the RTVE database, which has been used in the Albayzín evaluations. We employ a traditional system based on Hidden Markov Models with Gaussian Mixture Models. Results show that, although the task is difficult, in restricted conditions we obtain recognition results which determine that using eigenlips in combination with deep features is the best visual approach.

Index Terms: lipreading, machine learning, speech technologies, computer vision, hidden markov models, deep learning

1. Introduction

During a conversation, our brain is responsible for combining information obtained from multiple senses in order to improve our ability to understand the message we are perceiving. Different studies have shown the importance of presenting visual information in these situations, as well as its relationship with the sounds produced. Principally, we stand out the studies carried out by McGurk and McDonald [1], where they demonstrated that if the mouth expression does not match with the emitted sound, the listener was confused, perceiving a sound different from what it really was. Nevertheless, lipreading is a complex task whose objective is to interpret speech when audio is not available. By dispensing with a sense as crucial as hearing, since this signal presents a greater amount of information regarding speech recognition, it will be necessary to be aware of the challenge that this lack presents.

Our ideal purpose is to build a system capable of imitating the human ability to interpret continuous speech by reading the lips of the speaker. Due to the absence of acoustic cues, some of the main challenges we have to deal with are visual ambiguities and silence modelling [2, 3]. Therefore, an essential factor is to identify a suitable representation that manages to capture the nature of lip movements and how it affects to the recognition quality. Consequently, the central core of this paper deals with an analysis of speech visual features [4, 5].

For this comparison, a fixed decoding system must be chosen. In our case, we employed a traditional approach to define the automatic system, in other words, a system based on Hidden Markov Models combined with Gaussian Mixture Models (GMM-HMM), an approach that has been widely used in Acoustic Speech Recognition (ASR) [6]. Although this is not the state-of-the-art for speech-related signal recognition, it is an appropriate option for comparing the different possibilities for feature extraction. Unlike in ASR, when we deal with Visual Speech Recognition (VSR) our basic speech unit is not the phoneme, but the one known as the viseme, which is associated with the representation of the phoneme on the visual domain [7]. Unfortunately, there is not direct or one-to-one correspondence between them, which causes visual ambiguities. An interesting work [2] describes a phoneme-to-viseme mapping for Spanish and concludes its usefulness compared to recognition through phonemes directly. However, we decided to establish the phoneme like basic speech unit in our work as many authors have done [8, 9, 10].

Apart from that, in order to estimate our system, an audiovisual corpus focused on continuous speech has been built from a subset of the RTVE database [11], with Spanish being the language to be interpreted. The RTVE database is a well-known database which has been used in the Albayzin evaluations [12]. Taking into account all these aspects, we can integrate our task both in the field of Speech Technologies and Computer Vision.

In relation to the rest of the paper and its organization, we mention that Section 2 provides the context and historical evolution around the task of VSR or Automatic Lipreading Recognition (ALR). Section 3 presents several details regarding the built audiovisual corpus. Then, Section 4 describes the different visual approaches considered in order to represent the nature of lip movements. Section 5 shows the experimental process carried out in our work, as well as certain insights and comments regarding our results. Finally, conclusions and future lines of research are offered in Section 6.

2. State of the art

In its origins, automatic speech recognition systems focused only on acoustic information, since this signal is more informative to distinguish phonemes [13]. Nowadays, these models are powerful systems capable of understanding spoken language with great quality [14]. However, when the acoustic signal is damaged or corrupted, the performance of these systems decline considerably [13, 10]. Therefore, many authors have studied how the incorporation of visual cues alongside acoustic information cause a significant improvement over interpretations supplied by the system in these situations [10, 15]. Additionally, several studies related to Silent Speech Interfaces (SSI) [16] were carried out to deal with the possible absence of the acoustic signal in the field of Speech Technologies.

In this way, in the last decades there has been an increase in the interest of decoding speech using exclusively the informa-

10.21437/IberSPEECH.2021-47

tion from the visual channel. As Fernandez-Lopez and Sukno suggest in their review [13], advances achieved by this type of systems have been conditioned, among other reasons, by the evolution reflected over available audiovisual databases. These databases began by tackling simple tasks from alphabet and digit recognition, such as AVLetters [4] and CUAVE [17] corpora. More recent datasets provide the necessary support to estimate approaches in charge of interpreting spontaneous speech, for instance, *Lipreading Sentences in the Wild* (LRS) [18]. Regarding Spanish, we stand out the VLRF corpus [8], despite the fact that it differs from our objectives by recording the scenes under controlled conditions and ensuring that speakers strain theirselves to vocalize adequately and expressively. Lately, the CMU-MOSEAS database [19] has been compiled, among other languages, for Spanish. This is an interesting corpus, as it provides a multimodal point of view, supplying information related with the emotions and subjectivity expressed by the speaker.

At the beginning, these tasks were developed under a traditional paradigm, that is, mainly through the well-known HMMs. Since this is our case, we highlight some publications, such as studies carried out by Thangthai, Cox, and Howell, among others [20, 9], where they employed an HMM per phoneme, evaluating both dependent and context-independent models. On the other hand, in relation to Spanish, we mention again the paper developed by Fernandez-Lopez and Sukno [8], where we emphasize their study regarding recognition at the phoneme level over the VLRF corpus. However, the research has gravitated towards Deep Learning technologies. More concretely, end-to-end architectures, formed by combining Long Short Term Memory (LSTM) [21] and Convolutional Neural Networks (CNN) [22], have been the most widely used topology. In fact, Zisserman [18] reached the state-of-the-art in continuous VSR by employing this approach and achieving an error rate of around 50% at word level.

3. Audiovisual corpus

As we mentioned above, we have compiled an audiovisual corpus focused on the task of continuous lipreading recognition, where we could find a large number of speakers in a wide range of scenarios, including variations on intra-personal aspects or light conditions. We compiled it from a subset of the RTVE database [11] which has been employed in Albayzín evaluations [12]. The RTVE database is made up of different programs broadcasted by *Radio Televisión Española* but, in our work, we compiled the corpus only from the news program 20H broadcast by the *Canal 24 horas*. In this program, we have selected scenes where a unique speaker talks from different distances to the camera and in diverse scenarios, either inside a record studio or in outdoor locations. Furthermore, the speaker does not always maintain a frontal plane but can sometimes adopt tilted postures. Other details regarding the compiled dataset are shown in Table 1.

Table 1: *Details regarding the compiled audiovisual corpus.*

Language	Spanish		
Resolution	480×270 pixels, 30 frames/second		
Speakers	57	**Males:** 17	**Females:** 40
Duration	∼3 hours		
Utterances	2792		
Vocabulary	2885		
Running words	∼35k		
Phonemes	23		
Words per utterance	**Median:** 10	**Max:** 62	**Min:** 1
Phonemes per utterance	**Median:** 46	**Max:** 270	**Min:** 5

Before continuing, it is necessary to mention that the resolution of 480×270 pixels refers to the full record scene. Our region of interest, that is, the speaker's mouth, is of variable dimensions. Therefore, we stablished its size in 32×16 pixels.

4. Speech visual features

In the literature, a large number of approaches regarding the visual representation of speech have been studied. Many authors [4, 23, 24] have employed traditionals techniques to extract these features, such as Principal Component Analysis (PCA), Active Appearence Models (AAM), or Optical Flow. Moreover, other authors have delegated the responsibility of extracting visual features on neural networks [25, 26, 5], and more specifically on Convolutional Autoencoders.

However, there is no consensus or agreement in the literature on what is the best option for the extraction of visual speech features. Therefore, in our work we studied three types of features that we describe in subsections from 4.1 to 4.3. In all the approaches, the use of the OpenCV library [27] and the Dlib toolkit [28] allowed us to identify 68 facial landmarks. From some of these landmarks, as left part of Figure 1 reflects, we were able to extract our region of interest.

4.1. Geometric features

In this first approach, the study of continuous sign language carried out by Hermann Ney and other authors [29] is taken as a reference. In this way, we defined, thanks to the location of landmarks described above, a set of 18 high-level features, such as width, height, or area of speaker's mouth. However, when the speaker is more or less close to the camera the same metric (for example, mouth's width) would acquire a value in different magnitudes, even in the case of the same mouth posture or physiognomy. For this reason, we decided to locate a more stable region where each of the measured distances can be properly normalized. This is the region highlighted by a larger blue rectangle on the right side of Figure 1. Thus, the resulting geometric features are normalized using the size of this area.

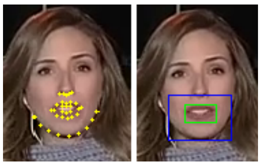

Figure 1: *Aspects regarding geometric features.*

4.2. Eigenlips

This concept is influenced by the studies carried out on facial recognition [30]. Then, as other authors did [3], after computing PCA over our training set we could obtain the eigenlips shown in Figure 2. The first component, as suggests the image, stands out the lip corners, since these are the parts that suffer the greatest deformation throughout a speech. As for the rest of eigenlips, some emphasize lip contours while others highlight zones where we can find teeth or tongue. These are aspects that we can not reach when we work with pure geometric features, but that are of vital importance for visual speech recognition.

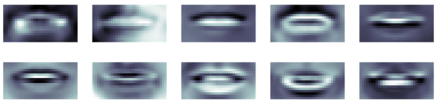

Figure 2: *Eigenlips obtained after applying PCA.*

Another important issue is that, with the intention of making easier the extraction of these and the features described in Subsection 4.3, we apply a mouth alignment. In other words, as Figure 3 suggests, we rotate those regions of interest where we observe that the speaker's mouth is tilted or inclined.

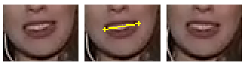

Figure 3: *Schematic process of mouth alignment.*

4.3. Deep features

The last approach, as we mentioned above, is based on Convolutional Autoencoders [31], a neural network whose main purpose consists of reconstructing the image received as input from an abstract and compact representation which has been obtained previously from the original image. Once this statistical model is trained, we can dispense with decoder because it is the encoder the component we need to extract our visual features. This scheme is reflected in Figure 4.

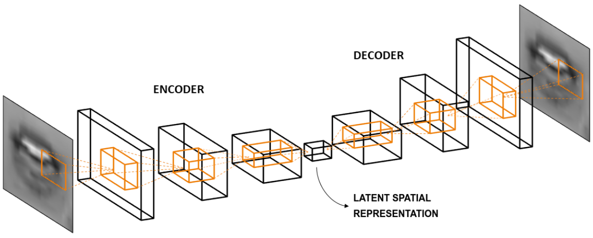

Figure 4: *Scheme of the proposed Convolutional Autoencoder.*

The encoder architecture is entirely based on that presented in [25], since they employ low-resolution images as is our case. Thus, we were able to obtain high quality reconstruction results, as shown in Figure 5. Finally, we remind that in these features we apply a mouth alignment too.

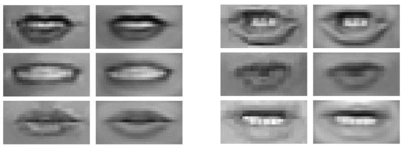

Figure 5: *Reconstructions examples obtained by the defined Convolutional Autoencoder. For each images block, left column: original image; right column: reconstruction.*

5. Experiments

All our experiments were performed with the Kaldi toolkit [32]. This toolkit provides the support to build high performance systems focused on Speech Technologies, from traditional approaches to hybrid models or systems based entirely on Deep Learning. In our case, as we stated in the introduction, we employed a traditional GMM-HMM system. This system is formed by three modules: the Optical Model, in charge of interpreting the pronounced phonemes from visual features sequence; the Lexical Model, responsible for building the words from the phonemes provided by the previous module; and the Language Model, capable of combining the words provided by the Lexical Model with the intention of generating the message interpreted by the system. On the other hand, our corpus is divided into two partitions, allocating to the test set those speakers who did not reach a minimum of seconds. Thus, we get a train set composed of 2672 utterances emitted by 43 different speakers, reaching around 3 hours of data, whereas the test partition comprises 120 samples from 13 speakers, covering 0.13 hours of utterances. Then, due to the limited amount of available data, we had to estimate a context-independent system, also know as monophonic system.

At the beginning, we carried out several speaker-independent experiments with an Open Language Model, but because of the difficulty of the task, we did not achieve minimally acceptable results (error rates greater than 90%). Consequently, it was necessary to make experiments in a more constrained scenario in order to obtain conclusions on the use of the different features. Therefore, we decided to relax the task complexity by employing a Closed Language Model. In other words, a Language Model estimated only from the text included in the test partition. In this way, the system has a reduced set of alternatives when it interprets the message, which allows us to focus our experiments on the performance of the different types of features. In fact, as suggested in [8], an acceptable recognition at phonemic level does not necessarily imply a good quality performance when word level decoding message is carried out.

Our first experiment focused on studying HMM's topology, one of the most relevant factors in relation to temporal alignment of visual data. The classic topology in ASR (3 states left-to-right with self-loops) provided us a poor recognition rate. Then, employing raw geometric features, we tested several topologies. After these experiments, we observed that if we reduce the number of states and we add transitions of each of these states to the final state, system performance obtains, in general, a considerable improvement. According to these results, we believe that this behaviour may be caused by the limited frequency of information (30 frames/second) that presents the visual channel with respect to the standard representation (Mel Frequency Cepstral Coefficients, MFCC) of acoustic data (100 frames/second) [6]. Consequently, in the rest of our work we employed the topology shown in Figure 6: 2 states left-to-right with self-loops and skip transitions.

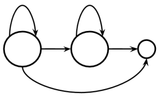

Figure 6: *HMM's topology employed in our experiments.*

Once this aspect is specified, we can address the analysis of the visual speech features with the intention of determining which of them, either isolated or combined with each other, is the best representation to capture the nature of lip movements. In addition, we study how the incorporation of temporal or delta-delta coefficients, used in ASR [6], affects to the recognition quality. More precisely, we study different contexts, both a greater and a lower scope. On the other hand, as Lan presented in [23], we have observed how applying a z-score normalization causes a considerable recognition quality improvement. Then, in our analysis we studied two types of normalization on all the features presented in Section 4:

- Normalization per speaker: all the utterances of a speaker are taken in the normalization; this is aimed at mitigating the differences that may exist in the aspect of the speaker in his/her different utterances (light conditions, facial hair, lipstick colour, temporal scars or marks in the mouth, ...).
- Normalization per utterance: the normalization is for each single utterance; this is done with the intention of

Table 2: *Results (WER) for visual speech features with speaker and utterance normalization.* ***raw****: refers to the raw features, without adding any type of temporal coefficient.* $\Delta\Delta_x$*: applies on the features the coefficients* delta-delta *with a context of x frames. Best result for each set of features and normalization in bold.*

	***z-score* normalization per speaker**				***z-score* normalization per utterance**			
Features	**raw**	$\Delta\Delta_1$	$\Delta\Delta_2$	$\Delta\Delta_3$	**raw**	$\Delta\Delta_1$	$\Delta\Delta_2$	$\Delta\Delta_3$
geometricFeats	51.3±8.5	49.2±8.2	**37.7±8.2**	50.5±7.8	46.1±8.8	49.8±8.6	**36.8±7.1**	48.7±8.0
eigenLips	71.6±8.6	60.6±8.5	**57.1±8.3**	65.8±7.5	66.6±8.8	**49.9±8.1**	55.6±8.5	53.0±8.4
deepFeats	46.6±8.5	**31.7±7.3**	35.9±7.7	39.1±7.8	72.4±7.9	58.9±8.2	54.8±8.9	**54.7±8.2**
geometricFeats+eigenLips	29.6±7.5	30.3±6.5	34.2±7.2	**26.6±6.0**	**26.1±6.2**	34.6±7.1	31.2±6.6	34.9±6.4
geometricFeats+deepFeats	32.9±7.9	**29.8±7.1**	34.1±7.0	41.3±6.7	**29.3±7.7**	36.5±7.0	33.0±6.9	41.3±7.6
eigenLips+deepFeats	45.2±8.2	33.6±7.9	**23.7±6.4**	35.4±7.1	38.0±7.5	29.6±6.8	**26.8±7.2**	31.0±7.2
geometricFeats+eigenLips+deepFeats	30.4±6.8	**27.3±6.5**	33.4±6.5	41.5±6.7	**34.6±7.7**	36.4±6.7	**34.6±6.8**	43.0±6.8

reducing the differences among the different utterances from different speakers (i.e., to obtain speaker independency) and conditions (e.g., focus variability).

All the results are evaluated by the well-known Word Error Rate (WER) with 95% confidence intervals obtained by the boostrap method described in [33].

Results are presented in Table 2. First, we stand out that the incorporation of delta-delta coefficients cause, in general, an improvement on system performance. Of course, depending on the type of features or normalization, it is convenient to use one context or other.

On the other hand, if we focus only on those experiments that study the visual features individually, we conclude that deep features are the best representation, in isolated way, to address VSR, as long as we normalize per speaker. In contrast, when we apply a normalization per utterance, we notice how the quality of these features suffer a drastic deterioration, while the rest of visual approaches improve their results. This may mean that deep features, as they are highly dependent on the pixel values, are more affected by changes in light conditions or certain intra-personal aspects. In this way, if these features are processed by a normalization per speaker, they are more benefited in order to interpret speech visually. In contrast, geometric features are more dependant on the location of the specific pixels, and thus utterance normalization fits better for these features. Eigenlips depend on both specific pixels and their values, which makes it to be the worst option when normalizing by speaker and to not improve geometric features when normalizing by utterance.

Regarding the experiments that explore the combinations of visual features, we confirm that, as a general rule, feature combination produces a decrease in error rates. Therefore, we can deduce that the studied features complement each other and manage to provide a more robust representation. Nevertheless, this is not always the case; in certain occasions these results are overlapped with the best error rates obtained in experiments where features were employed individually. On the other hand, the eigenlips and deep features combination, if we normalize per speaker, is established as the best approach to address the automatic lipreading, reaching around 23.7% WER, although differences are not significant with respect to only using deep features. Seemingly, thanks to appearance aspects contained in eigenlips and the great potential that deep features demonstrated regarding mouth physiognomy reconstruction, a high quality representation has been achieved. However, it is worth noting that if we do not incorporate delta-delta coefficients, we can verify how the geometric features and eigenlips combination improves the performance of the previously mentioned approach.

Finally, we stand out that the combination of all the features analyzed in our work forms a representation with a large dimension. This fact can cause difficulties when modelling data statistically. Consequently, unlike most cases in which isolated features are used, introducing temporal coefficients does not always imply better results.

6. Conclusions

In this paper, an extensive study has been carried out regarding visual speech features in order to address an automatic lipreading task for natural Spanish. Therefore, in addition to compiling a preliminary audiovisual corpus, approaches based on both traditional techniques and Deep Learning architectures have been addressed to represent the nature of lip movements. After our experiments, we conclude that the combination of eigenlips and deep features, as long as we apply certain aspects such as normalization and temporal coefficients, provide the best approach to interpret speech visually. On the other hand, aspects in relation to temporal alignment of visual data have been studied. More concretely, according to the results obtained in this work, the HMM's topology has been modified regarding standards in acoustic speech recognition.

However, visual speech recognition remains an open problem. Lipreading is a complex task where researchers have to face with visual ambiguities and other issues such as silence modelling. In addition, there is not a consensus or agreement regarding a suitable visual speech representation. For these and others reasons, further research is necessary. Authors such as Fernandez-Lopez and Sukno [13] have suggested that future lines of research should be developed around temporal alignments and context modelling. On the other hand, we consider to shift our study towards a pure Deep Learning approach. More precisely, we aim at an end-to-end architecture whose parameters, including those in charge of extracting visual features, are estimated according to the mistakes identified in the message decoding stage. In other words, we believe that this direct learning could be more useful than addressing the task with a traditional approach, where each module is independent of the other. In order to achieve this objective, we must increase the number of seconds which forms the compiled audiovisual corpus. In this way, we expect to get a large amount of data which represents the nature of natural speech and be able to estimate our statistical models appropriately. Finally, it would be interesting to study whether a suitable viseme-phoneme correspondence for Spanish can lead to advances in the matter.

7. Acknowledgements

This work was partially supported by Generalitat Valenciana under project DeepPattern (PROMETEO/2019/121) and by Ministerio de Ciencia under project MIRANDA-DocTIUM (RTI2018-095645-B-C22).

8. References

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

[2] A. Fernandez-Lopez and F. M. Sukno, “Optimizing phoneme-to-viseme mapping for continuous lip-reading in spanish,” in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2017, pp. 305–328.

[3] K. Thangthai, “Computer lipreading via hybrid deep neural network hidden markov models,” Ph.D. dissertation, University of East Anglia, 2018.

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

[5] D. Parekh, A. Gupta, S. Chhatpar, A. Yash, and M. Kulkarni, “Lip reading using convolutional auto encoders as feature extractor,” in *5th International Conference for Convergence in Technology (I2CT)*. IEEE, 2019, pp. 1–6.

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

[7] C. G. Fisher, “Confusions among visually perceived consonants,” *Journal of speech and hearing research*, vol. 11, no. 4, pp. 796–804, 1968.

[8] A. Fernandez-Lopez, O. Martinez, and F. M. Sukno, “Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database,” in *12th International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 208–215.

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

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

[11] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Zotano, and A. de Prada, “Rtve2018 database description,” *Vivolab and Corporación Radiotelevisión Española, Zaragoza, Spain*, 2018, [Online] Available: http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf.

[12] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, “Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media,” *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[13] A. Fernandez-Lopez and F. M. Sukno, “Survey on automatic lip-reading in the era of deep learning,” *Image and Vision Computing*, vol. 78, pp. 53–72, 2018.

[14] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[15] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[16] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, “Silent speech interfaces,” *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

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

[18] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[19] A. B. Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, “Moseas: A multimodal language dataset for spanish, portuguese, german and french,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1801–1812.

[20] D. Howell, S. Cox, and B. Theobald, “Visual units and confusion modelling for automatic lip-reading,” *Image and Vision Computing*, vol. 51, pp. 1–12, 2016.

[21] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: continual prediction with lstm,” in *Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, vol. 2, 1999, pp. 850–855 vol.2.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[23] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, and R. Bowden, “Improving visual features for lip-reading,” in *Auditory-Visual Speech Processing*, 2010, pp. 142–147.

[24] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. C. Azemin, and J. Gubbi, “Lip reading using optical flow and support vector machines,” in *3Rd international congress on image and signal processing*, vol. 1. IEEE, 2010, pp. 327–330.

[25] I. Fung and B. Mak, “End-to-end low-resource lip-reading with maxout cnn and lstm,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2511–2515.

[26] K. Paleček, “Extraction of features for lip-reading using autoencoders,” in *International Conference on Speech and Computer*. Springer, 2014, pp. 209–216.

[27] G. Bradski, “The opencv library,” *Dr Dobb’s J. Software Tools*, vol. 25, pp. 120–125, 2000.

[28] D. E. King, “Dlib-ml: A machine learning toolkit,” *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[29] O. Koller, J. Forster, and H. Ney, “Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers,” *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015.

[30] K. Delac, M. Grgic, and P. Liatsis, “Appearance-based statistical methods for face recognition,” in *Proceedings of the 47th International Symposium ELMAR focused on Multimedia Systems and Applications, Zadar, Croatia*, 2005, pp. 151–158.

[31] Y. Bengio, “Learning deep architectures for ai,” *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” in *IEEE workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[33] M. Bisani and H. Ney, “Bootstrap estimates for confidence intervals in asr performance evaluation,” in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. 409–412.