A Comprehensive Analysis of Human-centric Audio-Visual Learning in Speech: A survey

Anonymous ECCV submission

Paper ID xxxx

Abstract. Audio-visual learning, which capitalizes on the relation between auditory and visual modalities, has gained substantial scholarly attention in conjunction with the successful rise of deep learning techniques. Recently, researchers began to harness these dual modalities to enhance the performance of single-modality models or solve problems that were far-fetched from being solved within a single modality. In this paper, we present a survey of the different audio-visual models and tasks that are centered around human communications; and we delve into the different techniques, models, and datasets used in such tasks. Our comprehensive analysis categorizes human-centric audio-visual learning into three distinctive categories: audio-visual communication improving, audio-visual empowering, and audio-visual safeguarding. We hope it can offer the research community a comprehensive overview of human-centric audio-visual learning and the challenges it encounters.

Keywords: Multimodality, audio-visual learning, speech, deep learning, survey

1 Introduction

We, human beings, employ our multiple senses to engage with and understand our surroundings. In our ongoing journey of interacting with the world around us, spoken language serves as the predominant form of interpersonal communication [Anwar: needs citation?]. For many decades, it was believed that the auditory signal, coming through the ear, is the only needed signal for speech perception among hearing-capable people [Anwar: needs citation??]. That was until studying hearing-impaired people shed more light onto our false assumption and helped us understand more about our auditory-visual perception of speech [16] which is developed in our brain as early as 18 weeks old [24] and deteriorates at advancing age [17].

While studying deaf children, Numbers et al. [4] have pointed out that hearing-capable individuals also make use of the eye in the reception of speech. Hutton et al. [8], Hutton [9], Duffy et al. [11], and Siegenthaler et al. [15] all have demonstrated that when patients both listen through a hearing aid and see the talker speaking, their word-identification scores generally are much higher than what they can obtain by listening alone. Dodds et al. [12] and Ewertsen et al. [16] have reported similar effects on the sentence level.

To understand more about our bi-modal perception of speech, more studies were directed to measure the impact of noise on our perception. Sumby et .al [6] have found out that participants were able to recognize words under a low signal-to-noise (S/N) ratio of -30 dB when aided with visual factors (i.e. face or lip movement), despite that S/N of -18 dB is the lowest ratio at which speech can be detected under earphones [3]. O'Neil [5] measured the impact of noise on different parts of the speech (consonants, vowels, words, phrases). Neely [7] focused on quantifying the visual contribution in terms of the angle and distance, while others ([14], [16], [17], [19], [20]) compared the performance of audio-visual combined signal against audio-only and visual-only signals, and ([22], [46]) studied how much the face, mouth, and lips contribute to the visual perception.

Visual signal (i.e. lip movements, facial expressions, and body language) is a reservoir of invaluable information that has the potential to significantly enhance human communication, especially when auditory signals are subject to constraining conditions. The visual modality can serve as a complement to auditory cues, facilitating a more robust and comprehensible communication process as in speech recognition [?] and speech translation [116]. Furthermore, harnessing the power of visual information can achieve generative and non-generative capabilities that can transcend traditional auditory-based communication paradigms; generative capabilities like Facetalking [?] and lip-to-speech generation [?] and non-generative capabilities like audio-visual temporal synchronization (AVTS) [?] and speaker localization [?]. Additionally, integrating auditory and visual signals can also play a pivotal role in securing human communications as in DeepFake detection [?] and lip-based authentication [?].

Currently, the research community lacks a comprehensive analysis of the audio-visual learning that is concerned with human communication and we take the liberty to fill this gap. In our paper, we provide a thorough analysis of the human-centric audio-visual speech tasks and models, we categorize them into three distinct categories: 1) audio-visual communication improvement, 2) audio-visual empowerment, and 3) audio-visual safeguarding. Then, we provide full stats of the relevant audio-visual datasets and evaluation benchmarks. Finally, we end the paper by enumerating the challenges that encounter researchers in the field.

2 Related Work

Within the domain of multimodal machine learning and audio-visual learning, there exists a number of survey papers that cover a broad spectrum of areas. Several surveys are dedicated to establishing theoretical foundations and taxonomies for multimodal machine learning. Sun et al. [97], for instance, delve into the classical theories of multimodality, such as Canonical Correlation Analysis [2] and Co-learning [63].

In a similar fashion, Xu et al. [98] categorize multimodal learning paradigms into three distinct categories: Co-training, Multiple Kernel Learning, and Subspace Learning, while Baltrušaitis et al. [105] enlist the challenges that multimodal learning encounter and group them into five core categories: representation, translation/mapping, alignment, fusion, and co-learning. Other surveys, like Lahat et al. [101] and Atrey et al. [94], focused on data fusion, specifically why we need data fusion and how to perform it. Others, like Li et al. [106] and Zhang et al. [109], shift their focus to the mid-level representation.

Additionally, other surveys narrowed their focus to specific subcategories of multimodal learning. Mogadala et al. [112] concentrated on ten tasks that combine natural language processing with computer vision covering the models, methods, datasets, and evaluation measures. Others, such as Chai et al. [113], focused on more visual-related tasks such as Media Captioning, Visual Question Answering, Visual Content Description, Multimodal Machine Translation, Text–Image Retrieval and Text-Image Generation.

And one of the earliest examples of multimodal research is audio-visual speech recognition (AVSR) [85] where Potamianos et al. xxx. Others focused on different audio-visual tasks. Ngiam et al. [96] employed speech classification as a case study for multimodal learning using Restricted Boltzmann Machine (RBM) [87]. Michelsanti et al. [111] focused on two audio-visual-based tasks, AV Speech Enhancement (AV-SE) and AV Speech Separation (AV-SS). Also, they reviewed the different methods used to extract the features, fuse the audio-visual modalities, and the models' objective functions. Johnson et al. [118] focused on the interpretability side of audio-visual learning , Shivappa et al. [95] focused on audio-visual information fusion, and Ivanko et al. [117] focused on audio-visual speech recognition.

On the other hand, there were other surveys that were dedicated to audio-visual learning in general. Vilça et al. [114] review different models that were used in audio-visual learning, e.g. Transformers [104], AutoEncoders (AE) [122], and Generative Adversarial Network (GAN) [99]. Zhu et al. [110] were interested in audio-visual learning tasks and they grouped them into four different categories: Audio-visual Separation and Localization, Audio-visual Correspondence Learning, Audio-visual Generation, and Audio-visual Representation Learning.

Finally, Wei et al. [115] enlisted around 20 different audio-visual tasks and categorized them into three distinct categories: First, Audio-visual Boosting where one modality provides complementary information to the other modality leading to more robust models and boosting the performance. Second, Cross-modal Perception is where either one of the two modalities is absent and deduced through transferred knowledge or generated according to the information from the other one. Third, Audio-visual Collaboration is where the two modalities go hand-in-hand to perform the task.

In this paper, we fill in the gap and present a survey of the different audio-visual models and tasks that are more centered around human communications. In our comprehensive analysis, we categorize human-centric audio-visual learning into three distinctive categories: 1) **Communication Improvement** dedicated to all tasks that improve human communication, 2) **Empowerment** dedicated to the tasks that give human beings more capabilities, and **Safeguarding** dedicated for the tasks that are related to the human privacy and security.

3 Human-centric Audio-Visual Learning

better naming??? supplementary (modalities supporting each other) Vs complementary (modalities complete each other, i.e one modality can't work without the other.

3.1 AV Improving Communication

3.2 AV Empowerment

3.3 AV Safeguarding

xxx

3.3.1 Audio-visual Speech Recognition

In the late 1980s, automatic speech recognition (ASR) models started to get accurate results under optimal conditions [120]. However, there was a huge degradation in ASR performance under noisy environments. To overcome this, multiple techniques were introduced to either reduce or cancel the impact of noise on the speech signal, such as “Spectral Subtraction” [21] which estimates the noise envelope from the signal and subtracts it from the noisy speech and “Adaptive Filtering” [10] which adjusts itself continuously to minimize the residual interference affecting the target signal.

Inspired by the human's bimodal perception of speech [18] [24], audio-visual speech recognition (AVSR) was also considered as an alternative approach to enhance ASR performance [14], especially in noisy environments [6] [121] [27] [29]. As pointed out by [37], AVSR is the perception of speech by combining the auditory signal with the visual signal represented in lip movement.

One of the earliest works on AVSR was done in 1988 by Petajan et al. [30] where they combined acoustic and visual signals to perform audio-visual letter recognition on a vocabulary of 100 words consisting of English letters and digits. To perform this task, the proposed model incorporated dynamic time warping and vector quantization [25] on binary image sequence (black and white) of the oral-cavity of four speakers forming visemes [1] codebook. One year later, Yuhas et al. [33] (followed by [35]) created the first neural network to perform audio-visual vowel recognition. The neural network was very simple as it consisted of a Multi-layer Perceptron (MLP) containing five neurons and a sigmoid activation function to process a reduced area of interest (AoI) of 20×25 pixels centered around the mouth of static images.

In the 1990s, AVSR research started to gain a considerable amount of attention. In [36], authors used Time-Delay Neural Network (TDNN) [32] to process acoustic and visual features separately. Bregler et al. [38] [39] and Duchnowski et al. [43] [45] used Multi-state TDNN (MS-TDNN) networks, which is an expanded

[1] As defined in [13], a viseme or a “visual phoneme” as the smallest visibly distinguishing unit of a given language, similar to acoustic phoneme.

model of TDNN, and the model's acoustic and visual outputs are fused together by a weighted sum. Following that, a Dynamic Time Warping (DTW) layer [26] to find the optimal path of phone-hypotheses for the word models. Since the dimension for the visual signal is relatively large, [43] proposed using dimensionality reduction methods, such as Principal Components Analysis (PCA) [40], Linear Discriminant Analysis (LDA) [1], and Discrete Fourier Transform coefficients [34] without significant loss in performance.

Besides neural networks, Hidden Markov Models (HMM) were quickly adopted for AVSR due to their success in ASR [28] and their data efficiency. Additionally, some work took the best of both worlds creating a hybrid MLP-HMM model [119] [42]. In HMM, transition probabilities are encoded in links between nodes representing phonemic segments. Silsbee et al. [41] [54] used two HMM models to represent both modalities and a weighted sum to combine both HMM outputs in order to make a decision. Adjoudani et al. [44] [47] compared fusing both modalities on the feature level (i.e "Direct Identification Model") similar to [43] [49] [60] [62] [65]; and on the decision level (i.e "Separated Identification Model) similar to [51] [52] [38] [57] [58] [59] [61] [72]. Rogozan et al. in their comparison [64] used a hybrid of both. Additionally, Teissier et al. [66] pointed out two more fusion methods that aren't as common. Namely, "Dominant Recoding" [33] [35] [53] where the visual input is recoded into the representation of the auditory modality which is the dominant signal, and "Motor Recoding" [53] where both inputs are projected into an amodal (neither auditory nor visual) common space and fused in that space.

Due to the better performance of HMMs over neural networks, many researchers adopted different HMM variants, such as cross-product HMM [55] [67] [81], coupled HMM [82] [83] [69], Factorial HMM [82], and multi-stream HMM (MS-HMM) [55] [75] [78] [84]. MS-HMM was originally proposed for multi-band audio-only speech recognition [48], it linearly combines the class log-likelihoods based on the audio- and visual-only observations on the HMM state-level using both independent [68] [67], and joint training schemes [68] [79]. By default, MS-HMM assumes a complete synchrony between the two streams which is consistent with the Fuzzy Logical Model of Perception (FLMP) [31] which hypothesizes that signals from the eye and ear pass through several independent stages of neural processing before they come together in the parietal region of the brain.

However, later studies [42] [100] pointed out that the visual signal and the audio signal are correlated but not synchronous. [42] found out that a temporal shift by 120 ms achieves maximum mutual information between audio and visual signals, which means that on average the visual signal precedes the acoustic signal by 120 ms. A more elaborate study [100] came after and stated that the audio-visual asynchrony fluctuates, varying between 20 ms audio lead to 70 ms audio lag; and up to 200 ms audio lag for more complex speech structures. This has led to the need for asynchronous models capable of dealing with this issue, such as multi-stream state asynchronous HMM models [73] [71] [80] and multi-state asynchrony Dynamic Bayesian Network [89] [90] [93].

By the 2000s and with the rise of these advanced multi-stream models, more research was directed towards answering two very important questions:

1. *How to best represent audio and visual signals?*
2. *How to balance the audio/visual representations when integrating?*

To answer the first question, various studies explored different representational methods for audio and visual signals. For audio signal, Mel-frequency Cepstrum Coefficients (MFCC) [23] was quickly considered the status quo. However, different methods, such as Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) [42] and Discrete Wavelet Coefficients (DWC) [72], were considered but with less impact. For visual representations, a great amount of effort has been spent on engineering different methods, which can be broadly categorized into three categories: "image-based" (also known as "bottom-up"), "model-based" (also known as "top-down"), and a combination of both. In "image-based" approaches, compression algorithms were used to directly estimate low-level visual features from the image, such as discrete cosine transform [74] [77], PCA [40], and LDA [1]. In "model-based" approaches, a priori lip/mouth shape representation is embedded in a model that extracts higher-level (mainly lip contours, tongue, and teeth positions) and the model's parameters are used as visual features. Two methods were the most common approaches in this category, active shape models (ASMs) [50] and active appearance models (AAMs) [70] [76] [88]. "Image-based" approaches are advantageous since they do not require dedicated lip-shape models or hand-labeled data for training. However, they are vulnerable to any variance in lighting conditions, shifting, and rotations of input images. That's why a hybrid of these two methods was adopted as in [56] [68] [78].

To answer the second question, different stream weighting schemes were .

With the increasing availability of data and computing resources, the paradigm of feature engineering. This has allowed researchers to switch focus from finding good visual features to designing neural architectures for learning such representations.

By the start of the 2010s decade, more large high-quality data [?] [?] became accessible to the research community and that led to the raise of neural network based models such as Deep Belief Network, RNN, and LSTM

Application of deep learning to multi-modal analyses was presented in [19] which describes multi-modal, cross-modal and shared representation learning and their applications to AV-ASR. In [12], Deep Belief Networks(DBN) are explored. In [18] the authors train separate networks for audio and visual inputs and fuse the final layers of two networks, and then build a third DNN with the fused features. In addition, [18] presents a new DNN architecture with a bilinear soft-max layer which further improves the performance. In [20] a deep de-noising auto-encoder is used to learn noise robust speech features. The auto-encoder is trained with MFCC features of noisy speech as input and reconstructs clean features. The outputs of final layer of the auto-encoder are used as audio features. A CNN is trained with images from the mouth region as input and phoneme labels as output. The final layers of the two networks are then combined to train a multi-stream HMM.

Silsbee et al. [l] utilized vector quantization (VQ) of acoustic and visual data for their HMM based audio and video subsystems. Teissier et al. [8] utilized 20 FFT based 1-bark wide channels between 0 and 5 Khz for acoustic features and inner lip horizontal width, inner lip vertical height and inner lip area for the visual features. Chiou et al. [9] utilized active contour modeling to extract visual features of geometric space, the Karhunen-Lokve transform (KLT) to extract principal components in the color eigenspace, and HMMs to recognize the combined video only feature sequences. Potamianos et al. [3] used Fourier descriptor magnitudes for a number of Fourier coefficients, width, height, area, central moments, and normalized moments as contour features, and image transform features. It is worth noting here that the early visual feature extraction techniques are not affine (translation, rotation, scaling, and shear) invariant.

Various survey papers were interested in revewing different parts of the AVSR, such as [?] which was interested in overviewig AVSR models, and [?] which was interested in the different fusion methods, ...etc.

Task	Auditory	Visual	Model	Dataset	related papers
Vowel Recognition					
Phoneme Recognition					
LVCSR					

Table 1. Caption

3.3.2 Audio-visual Speech Translation

In this part, we are going to discuss "Audio-visual Speech Translation" (AVST) which is the task of translating speech signals in a source language to text in a target language with the aid of visual signals in the same source language (i.e face/mouth movement) which can enhance the model's robustness especially in noisy environments. AVST is different than "Audio-visual Translation" (AVT) which is the interlingual transfer of verbal language when it is transmitted and accessed both visually and acoustically [92], i.e translating products for the screen such as movies, TV series, plays, operas, ... etc that might include a textual description of the visual scene or an action demonstrated in the clip. AVST is a subset of "Multimodal Machine Translation" [108], which is the task of translating a sentence paired from the source language with an additional modality (e.g. audio modality in spoken language translation [91] or visual modality in image-guided [86] and video-guided translation [107]) into a different target language [102].

Despite AVST being the twin task for AVSR, the former task didn't find the same attention as the latter.

A speech translation system has been studied mainly for verbal information. However, both verbal and non-verbal information is indispensable for natural human communication.

AVST didn't find the same attention as AVSR despite the similarity due to the scarcity of high-quality data.

Also, there

3.3.3 Audio-visual Affect (Emotion) Recognition

Audio-visual Affect Recognition consists of laughter recognition, cry recognition,

4 Datasets

Originally, audio-visual datasets collected for research usually ended up in-house and very few made to available online freely. The large majority of the publicly available AV datasets were limited in size and scope, consisting mostly of tens of hours of spoken digits or command-and-control word sequences with a limited number of speakers recorded in a laboratory environment under controlled lighting conditions and face orientation.

4.1 Evaluation Benchmarks

5 Challenges

non-cleft lip, facial hear?, privacy-related issues, Distance

McGurk Effect: Also, Our audio-visual speech perception suffers from a bias well known as "The McGurk effect" [19], which stems from an inference between auditory signal and visual signal in the brain where subjects heard the syllable "ba" while watching a silent video of a person saying "ga", they eventually perceived the speech as a third sound: "da" [20]. And when these subjects listened to the auditory signal only, or when they watched the untreated video, they reported the syllables accurately. This phenomenon poses a potential challenge to audio-visual speech models. To the best of our knowledge, there has been no research work that has systematically examined the impact of this phenomenon on audio-visual artificial intelligence models.

Explainable AI: Interpretability, or explainable Artificial Intelligence (XAI) [103], is a very important aspect that we need to sincerely consider when dealing with human-level AI systems. Most of the published audio-visual models still lack the ability to illustrate why a model operates in a particular manner and under what conditions it may fail. While noteworthy work has already been made [118], there remains a substantial need for further research aimed at providing comprehensive explanations for the decision-making processes of such models. This is essential to increase the overall trustworthiness of audio-visual models.

6 Conclusion

References

1. Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.
2. Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.
3. George A Miller. The masking of speech. *Psychological bulletin*, 44(2):105, 1947.
4. Mary E Numbers and CLARENCE V Hudgins. Speech perception in present day education for deaf children. *Volta Rev*, 50:449–456, 1948.
5. John J O'Neill. Contributions of the visual components of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders*, 19(4):429–439, 1954.
6. William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.
7. Keith K Neely. Effect of visual factors on the intelligibility of speech. *The Journal of the Acoustical Society of America*, 28(6):1275–1277, 1956.
8. Charles Hutton, E Thayer Curry, and Mary Beth Armstrong. Semi-diagnostic test materials for aural rehabilitation. *Journal of Speech and Hearing Disorders*, 24(4):319–329, 1959.
9. Charles Hutton. A diagnostic approach to combined techniques in aural rehabilitation. *Journal of Speech and Hearing Disorders*, 25(3):267–272, 1960.
10. B Widrow. Adaptive filters 1: Fundamentals,‖ stanford electronics lab, 1966.
11. John K Duffy. Audio-visual speech audiometry and a new audio and audio-visual speech perception index. *Maico Audiological Library Series*, 5(9), 1967.
12. Elizabeth Dodds and Earl Harford. Application of a lipreading test in a hearing aid evaluation. *Journal of Speech and Hearing Disorders*, 33(2):167–173, 1968.
13. Cletus G Fisher. Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11(4):796–804, 1968.
14. Norman P Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of speech and hearing research*, 12(2):423–425, 1969.
15. Bruce M Siegenthaler and Vera Gruber. Combining vision and audition for speech reception. *Journal of Speech and Hearing Disorders*, 34(1):58–60, 1969.
16. HW Ewertsen, H Birk Nielsen, and S Scott Nielsen. Audio—visual speech perception a preliminary report. *Acta Oto-Laryngologica*, 69(sup263):229–230, 1970.
17. HW Ewertsen and H Birk Nielsen. A comparative analysis of the audiovisual, auditive and visual perception of speech. *Acta oto-laryngologica*, 72(1-6):201–205, 1971.
18. Norman P Erber. Auditory-visual perception of speech. *Journal of speech and hearing disorders*, 40(4):481–492, 1975.
19. Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.
20. John MacDonald and Harry McGurk. Visual influences on speech perception processes. *Perception & psychophysics*, 24(3):253–257, 1978.
21. Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120, 1979.
22. Quentin Summerfield. Use of visual information for phonetic perception. *Phonetica*, 36(4-5):314–331, 1979.
23. Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

24. Patricia K Kuhl and Andrew N Meltzoff. The bimodal perception of speech in infancy. *Science*, 218(4577):1138–1141, 1982.
25. Allen Gersho and Vladimir Cuperman. Vector quantization: A pattern-matching technique for speech coding. *IEEE Communications magazine*, 21(9):15–21, 1983.
26. Hermann Ney. The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):263–271, 1984.
27. N Michael Brooke and Eric D Petajan. Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics. In *Proceedings of the International Conference on Speech Input/Output: Techniques and Applications*, pages 104–109, 1986.
28. DB Paul, RP Lippmann, Y Chen, and CJ Weinstein. Robust hmm-based techniques for recognition of speech produced under stress and in noise. In *Speech Tech*, volume 86, pages 28–30, 1986.
29. Kathleen Ellen Finn. An investigation of visible lip information to be used in automated speech recognition. 1987.
30. Eric Petajan, Bradford Bischoff, David Bodoff, and N Michael Brooke. An improved automatic lipreading system to enhance speech recognition. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 19–25, 1988.
31. Dominic W. Massaro. Speech perception by ear and eye: A paradigm for psychological inquiry. 1989.
32. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989.
33. Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71, 1989.
34. Jae S Lim. Two-dimensional signal and image processing. *Englewood Cliffs*, 1990.
35. Ben P Yuhas, Moise H Goldstein, Terence J Sejnowski, and Robert E Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, 78(10):1658–1668, 1990.
36. David G Stork, Greg Wolff, and Earl Levine. Neural network lipreading system for improved speech recognition. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 2, pages 289–295. IEEE, 1992.
37. Quentin Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):71–78, 1992.
38. Christoph Bregler, Hermann Hild, Stefan Manke, and Alex Waibel. Improving connected letter recognition by lipreading. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 557–560. Ieee, 1993.
39. Christoph Bregler, Stefan Manke, Hermann Hild, and Alex Waibel. Bimodal sensor integration on the example of'speechreading'. In *IEEE International Conference on Neural Networks*, pages 667–671. IEEE, 1993.
40. Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.
41. Peter L Silsbee and Alan Conrad Bovik. Audio-visual speech recognition for a vowel discrimination task. In *Visual Communications and Image Processing'93*, volume 2094, pages 84–95. SPIE, 1993.

42. Christoph Bregler and Yochai Konig. " eigenlips" for robust speech recognition. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages II–669. IEEE, 1994.
43. Paul Duchnowski, Uwe Meier, and Alex Waibel. See me, hear me: integrating automatic speech recognition and lip-reading. In *ICSLP*, volume 94, pages 547–550. Citeseer, 1994.
44. Ali Adjoudani and Christian Benoit. Audio-visual speech recognition compared across two architectures. In *Eurospeech*, volume 95, pages 1563–1566, 1995.
45. Paul Duchnowski, Martin Hunke, Dietrich Busching, Uwe Meier, and Alex Waibel. Toward movement-invariant automatic lip-reading and speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 109–112. IEEE, 1995.
46. Bertrand Le Goff, Thierry Guiard-Marigny, and Christian Benoît. Read my lips... and my jaw! how intelligible are the components of a speaker's face? In *EUROSPEECH*, 1995.
47. Ali Adjoudani and Christian Benoît. On the integration of auditory and visual parameters in an hmm-based asr. *Speechreading by humans and machines: Models, systems, and applications*, pages 461–471, 1996.
48. Hervé Bourlard and Stéphane Dupont. A mew asr approach based on independent processing and recombination of partial frequency bands. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 1, pages 426–429. IEEE, 1996.
49. Barney Dalton, Robert Kaucic, and Andrew Blake. Automatic speechreading using dynamic contours. *Speechreading by Humans and Machines: Models, Systems, and Applications*, pages 373–382, 1996.
50. Juergen Luettin, Neil A Thacker, and Steve W Beet. Visual speech recognition using active shape models and hidden markov models. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 817–820. IEEE, 1996.
51. Uwe Meier, Wolfgang Hurst, and Paul Duchnowski. Adaptive bimodal sensor fusion for automatic speechreading. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 833–836. IEEE, 1996.
52. Javier R Movellan and George Chadderdon. Channel separability in the audio-visual integration of speech: A bayesian approach. *Speechreading by humans and machines: Models, systems, and applications*, pages 473–487, 1996.
53. Jordi Robert-Ribes, Michel Piquemal, Jean-Luc Schwartz, and Pierre Escudier. Exploiting sensor fusion architectures and stimuli complementarity in av speech recognition. In *Speechreading by Humans and Machines: Models, Systems, and Applications*, pages 193–210. Springer, 1996.
54. Peter Livingston Silsbee and Alan C Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, 1996.
55. Michael J Tomlinson, Martin J Russell, and NM Brooke. Integrating audio and visual information to provide highly robust speech recognition. In *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, volume 2, pages 821–824. IEEE, 1996.
56. Greg I Chiou and Jenq-Neng Hwang. Lipreading from color video. *IEEE Transactions on Image Processing*, 6(8):1192–1195, 1997.

57. Stephen Cox, Iain Matthews, and Andrew Bangham. Combining noise compensation with visual information in speech recognition. In *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*, 1997.
58. Thomas S Huang, Christopher P Hess, Hao Pan, and Zhi-Pei Liang. A neuronet approach to information fusion. In *Proceedings of First Signal Processing Society Workshop on Multimedia Signal Processing*, pages 45–50. IEEE, 1997.
59. Pierre Jourlin. Word-dependent acoustic-labial weights in hmm-based speech recognition. In *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*, 1997.
60. G. Krone, B. Talk, A. Wichert, and G. Palm. Neural architectures for sensor fusion in speech recognition. In *Proc. Auditory-Visual Speech Processing*, pages 57–60, 1997.
61. Gabi Krone, B Talk, Andreas Wichert, and Günther Palm. Neural architectures for sensorfusion in speechrecognition. In *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*, 1997.
62. Gerasimos Potamianos, Eric Cosatto, Hans Peter Graf, and David B Roe. Speaker independent audio-visual database for bimodal asr. In *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*, 1997.
63. Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.
64. Alexandrina Rogozan and Paul Deléglise. Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Communication*, 26(1-2):149–161, 1998.
65. Pascal Teissier, Anne Guerin-Dugue, and Jean-Luc Schwartz. Models for audio-visual fusion in a noisy-vowel recognition task. *Journal of VLSI signal processing systems for signal, image and video technology*, 20(1-2):25–44, 1998.
66. Pascal Teissier, Jordi Robert-Ribes, J-L Schwartz, and Anne Guérin-Dugué. Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing*, 7(6):629–642, 1999.
67. Stéphane Dupont and Juergen Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE transactions on multimedia*, 2(3):141–151, 2000.
68. Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, and Azad Mashari. Audio visual speech recognition. 2000.
69. You Zhang, Stephen Levinson, and Thomas Huang. Speaker independent audio-visual speech recognition. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 2, pages 1073–1076. IEEE, 2000.
70. Timothy F. Cootes, Gareth J. Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.
71. Hervé Glotin, D Vergyr, Chalapathy Neti, Gerasimos Potamianos, and Juergen Luettin. Weighting schemes for audio-visual fusion in speech recognition. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 173–176. IEEE, 2001.
72. Sabri Gurbuz, Zekeriya Tufekci, Eric Patterson, and John N Gowdy. Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 177–180. IEEE, 2001.

73. Juergen Luettin, Gerasimos Potamianos, and Chalapathy Neti. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 169–172. IEEE, 2001.
74. Iain Matthews, Gerasimos Potamianos, Chalapathy Neti, and Juergen Luettin. A comparison of model and transform-based visual features for audio-visual lvcsr. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 210–210. IEEE Computer Society, 2001.
75. Satoshi Nakamura. Fusion of audio-visual information for integrated speech processing. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 127–143. Springer, 2001.
76. Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Hervé Glotin, and Dimitra Vergyri. Large-vocabulary audio-visual speech recognition: A summary of the johns hopkins summer 2000 workshop. In *2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564)*, pages 619–624. IEEE, 2001.
77. Patricia Scanlon and R Reilly. Feature analysis for automatic speechreading. In *2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564)*, pages 625–630. IEEE, 2001.
78. Petar S Aleksic, Jay J Williams, Zhilin Wu, and Aggelos K Katsaggelos. Audio-visual speech recognition using mpeg-4 compliant visual features. *EURASIP Journal on Advances in Signal Processing*, 2002:1–15, 2002.
79. Guillaume Gravier, Scott Axelrod, Gerasimos Potamianos, and Chalapathy Neti. Maximum entropy and mce based hmm stream weight estimation for audio-visual asr. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–853. IEEE, 2002.
80. Guillaume Gravier, Gerasimos Potamianos, and Chalapathy Neti. Asynchrony modeling for audio-visual speech recognition. In *Proc. Human Language Technology Conference*, pages 24–27. Citeseer, 2002.
81. Sabri Gurbuz, Zekeriya Tufekci, Eric Patterson, and John N Gowdy. Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–2021. IEEE, 2002.
82. Ara V Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002:1–15, 2002.
83. Xiaozheng Zhang, Russell M Mersereau, and Mark Clements. Bimodal fusion in audio-visual speech recognition. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I. IEEE, 2002.
84. Etienne Marcheret, Stephen M Chu, Vaibhava Goel, and Gerasimos Potamianos. Efficient likelihood computation in multi-stream hmm based audio-visual speech recognition. In *Eighth International Conference on Spoken Language Processing*, 2004.
85. Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004.
86. Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2, 2006.
87. Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

88. Vassilis Pitsikalis, Athanassios Katsamanis, George Papandreou, and Petros Maragos. Adaptive multimodal fusion by uncertainty compensation. In *INTERSPEECH*, 2006.
89. Guoyun Lv, Dongmei Jiang, Rongchun Zhao, and Yunshu Hou. Multi-stream asynchrony modeling for audio-visual speech recognition. In *Ninth IEEE International Symposium on Multimedia (ISM 2007)*, pages 37–44. IEEE, 2007.
90. Louis Terry and Aggelos K Katsaggelos. A phone-viseme dynamic bayesian network for audio-visual automatic speech recognition. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.
91. Alex Waibel and Christian Fugen. Spoken language translation. *IEEE Signal Processing Magazine*, 25(3):70–79, 2008.
92. Delia Chiaro et al. Issues in audiovisual translation. *The Routledge companion to translation studies*, 141:165, 2009.
93. George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):423–435, 2009.
94. Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379, 2010.
95. Shankar T Shivappa, Mohan Manubhai Trivedi, and Bhaskar D Rao. Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, 98(10):1692–1715, 2010.
96. Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.
97. Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23:2031–2038, 2013.
98. Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.
99. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.
100. Jean-Luc Schwartz and Christophe Savariaux. No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology*, 10(7):e1003743, 2014.
101. Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.
102. Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016.
103. David Gunning. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, 2(2):1, 2017.
104. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
105. Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

106. Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018.
107. Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.
108. Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. Multimodal machine translation through visuals and speech. *Machine Translation*, 34:97–147, 2020.
109. Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.
110. Hao Zhu, Mandi Luo, Rui Wang, Aihua Zheng, and Ran He. Deep audio-visual learning: A survey, 2020.
111. Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation, 2021.
112. Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, 71:1183–1317, aug 2021.
113. Wenhao Chai and Gaoang Wang. Deep vision multimodal learning: Methodology, benchmark, and trend. *Applied Sciences*, 2022.
114. Luís Vilaça, Yi Yu, and Paula Viana. Recent advances and challenges in deep audio-visual correlation learning, 2022.
115. Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective, 2022.
116. Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *arXiv preprint arXiv:2303.00628*, 2023.
117. Denis Ivanko, Dmitry Ryumin, and Alexey Karpov. A review of recent advances on deep learning methods for audio-visual speech recognition. *Mathematics*, 11(12):2665, 2023.
118. David S Johnson, Olya Hakobyan, and Hanna Drimalla. Towards interpretability in audio and visual affective machine learning: A review. *arXiv preprint arXiv:2306.08933*, 2023.
119. Herve A Bourlard and Nelson Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 1994.
120. Kai-Fu Lee. *Automatic speech recognition: the development of the SPHINX system*, volume 62. Springer Science & Business Media, 1988.
121. Eric David Petajan. *Automatic lipreading to enhance speech recognition (speech reading)*. University of Illinois at Urbana-Champaign, 1984.
122. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation*, page 318–362. MIT Press, Cambridge, MA, USA, 1986.