PIAVE: A Pose-Invariant Audio-Visual Speaker Extraction Network

Qinghua Liu[1,2], Meng Ge[2,3], Zhizheng Wu[2], Haizhou Li[1,2,3]*

[1]Shenzhen Research Institute of Big Data, Shenzhen, China
[2]School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
[3] Department of Electrical and Computer Engineering, National University of Singapore, Singapore

{liuqinghua,gemeng,wuzhizheng,haizhouli}@cuhk.edu.cn

Abstract

It is common in everyday spoken communication that we look at the turning head of a talker to listen to his/her voice. Humans see the talker to listen better, so do machines. However, previous studies on audio-visual speaker extraction have not effectively handled the varying talking face. This paper studies how to take full advantage of the varying talking face. We propose a Pose-Invariant Audio-Visual Speaker Extraction Network (PIAVE) that incorporates an additional pose-invariant view to improve audio-visual speaker extraction. Specifically, we generate the pose-invariant view from each original pose orientation, which enables the model to receive a consistent frontal view of the talker regardless of his/her head pose, therefore, forming a multi-view visual input for the speaker. Experiments on the multi-view MEAD and in-the-wild LRS3 dataset demonstrate that PIAVE outperforms the state-of-the-art and is more robust to pose variations.

Index Terms: speaker extraction, multi-modality, pose variation problem, pose-invariant view

1. Introduction

The human brain has a remarkable ability to focus auditory attention on a particular voice by masking out the acoustic background in the presence of multiple speakers and background noises [1], that is called cocktail party effect [2]. With the advent of deep learning [3], neural approaches become increasingly popular in solving the cocktail party problem. The speaker extraction is one of such solutions that extracts the target speaker of interest from a multi-talk environment. It relies on an auxiliary cue to direct the attention towards the target speaker. The auxiliary cue may take different forms, such as pre-recorded reference speech [4, 5, 6], speech-synchronized video clips [7, 8, 9, 10], and the target speaker's spatial location [11, 12]. Notably, visual information is highlighted as a particularly useful cue for its immunity to acoustic noise and competing speakers [13] and is more informative than an audio cue [14].

Although recent audio-visual speaker extraction (AVSE) systems have demonstrated significant improvements in performance across a variety of standard datasets, the impact of head pose variations has not been studied. Head pose variations [15] as shown in Figure 1-(a) will result in changes in the orientation of the speaker's face in the camera view. Such variations adversely impact visually-assisted speech processing systems such as lipreading [16] and audio-visual speaker extraction.

We argue that an invariant pose could help audio-visual speaker extraction. Hence, we propose a Pose-Invariant Audio-

---

[*]Corresponding author

Figure 1: *(a) An illustration of the pose variation in daily conversations. (b) Pose-invariant faces corresponding to the above row of faces.*

Visual Speaker Extraction Network, namely PIAVE. This approach involves generating pose-invariant view (front view, in this paper) faces from the original face track, enabling that the model receives a consistent frontal view of the talker regardless of head pose variations. Furthermore, the model benefits from multi-view observation of the talking face, namely the generated pose-invariant view and the original pose orientation, thereby aiding in the more accurate identification of the target speaker and obtaining better performance in speaker extraction.

The contributions of this paper can be summarized as follows. Firstly, PIAVE represents the first step towards addressing the pose variation problem in AVSE. Secondly, we effectively generate the corresponding pose-invariant face for any given face image input, as shown in Figure 1-(b), which ensures stable visual input for the model. Lastly, when only one camera is used, PIAVE benefits from the multi-view observation of the target speaker, outperforming the state-of-the-art.

2. Speaker Extraction with Visual Cues

Figure 2 is an illustration of the workflow of the proposed PIAVE network during run-time inference. The network takes an audio mixture x_t and a video sequence of the target speaker's faces v_t as inputs. It could be described as:

$$\hat{y} = \text{Decoder}^a(\text{Separator}(a_e, v_e) \odot a_e), \qquad (1)$$

where $\hat{y}$ is the predicted target audio, $\odot$ is an operation for element-wise multiplication, a_e and v_e is encoded representations of the audio mixture and video sequence. The network consists of four parts: encoder, separator, decoder and pose normalizer. The pose normalizer will be described in Section 3.

The audio encoder converts the audio mixture into a spectrum-like representation in the time domain, and the audio decoder is used to transform the masked encoder representation to the target audio following [17].

The visual encoder extracts visual embeddings, which model the temporal synchronization and interaction between visemes and speech. It has a 3D convolutional layer followed
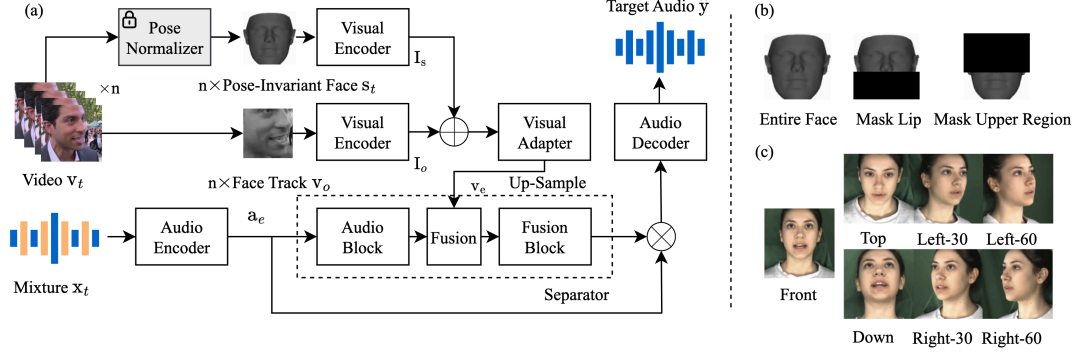
Figure 2: *(a) The overall architecture of the proposed PIAVE network. (b) An illustration of masking regions in Table 1 and 2. (c) Multi-view illustration of the MEAD dataset.*

by an 18-layer ResNet [18], which is pretrained on the lipreading task following [19]. It takes a sequence of pose-invariant faces and the original face track as input and outputs the fixed dimensional feature vectors $I_s \in \mathbb{R}^{d \times n}$ and $I_o \in \mathbb{R}^{d \times n}$ for the two video streams, where d denotes the feature dimension of the visual embedding and n denotes the number of frames in each video stream. These feature vectors are fused together using addition and passed through a visual adapter that models the temporal dependencies across each frame to capture the temporal dynamics of the visual input. Moreover, since the time resolution of the video stream and the audio stream is different, we upsample the visual embeddings along the temporal dimension to synchronize the audio stream and the video stream by nearest neighbor interpolation.

The separator is designed for estimating a mask to let pass the target speaker, and filter out others, conditioned on the encoded audio and visual features. To capture the long-term dependencies in the encoded audio features, the audio block employs a TCN network consisting of multiple TCN blocks, each with a depth-wise separable convolution, PReLU activation, and layer normalization operation. The output of the audio block is gathered with the encoded visual features v_e through concatenation and then we use two repeats of the TCN network to process the fused audio-visual feature. A convolutional layer is followed by the fusion block to generate the mask of the target speaker. It is then multiplied by the encoded audio representations of the mixture a_e to obtain the target audio $\hat{y}$.

During training, the Scale-Invariant Signal-to-Distortion ratio (SI-SDR) [20] between the predicted target audio $\hat{y}$ and the ground-truth target audio y is used to optimize the network from end to end:

$$\mathcal{L}_{\text{SI-SDR}} = -\rho(\hat{y}, y), \tag{2}$$

$$\rho(\hat{y}, y) = 20 \log_{10} \frac{||(\hat{y}^T y/y^T y) \cdot y||}{||(\hat{y}^T y/y^T y) \cdot y - \hat{y}||}. \tag{3}$$

3. Head Pose Normalization

3.1. Problem formulation

The variation of head poses has always been a challenge in both visual-only tasks and multi-modal processing. With large head pose variations, the intra-person variance of head representation may drastically increase, sometimes even exceeding inter-person variance. Recent research has shown that pose normalization consistently boosts the performance of face recognition [21, 22]. It should be noted that expression-free pose nor-

malization is required for face recognition tasks. However, in AVSE, lip movements and facial expressions that contribute to speech production correspond with phonetic content and have a strong impact on the ability of humans to focus their auditory attention [23, 24]. For this reason, expression-preserving pose normalization is required in AVSE.

In this paper, we propose pretraining the pose normalization module (i.e. pose normalizer in Figure 2-(a)) on the 3D face alignment and reconstruction task and keeping it frozen during the speaker extraction model training. The pose normalizer is expected to generate the expression-preserving pose-invariant face for any given face image input. The challenge is that large pose diversity makes it hard to distinguish facial expression variations and increases the modeling difficulty. To address this problem, we draw inspiration from [25] and propose to disentangle the head pose and facial expression to reduce the complexity of the problem and make it more tractable. In the following sections, we will introduce the pretraining setup for this dual face regression network and the approach we have adopted for head pose normalization.

3.2. Facial geometry representation

We separate the 3D face geometry into pose, mean shape, and deformation as:

$$G = f \times R \times S + t, \tag{4}$$

$$S = (\overline{S} + D). \tag{5}$$

where $G \in \mathbb{R}^{3 \times m}$ is the 3D mesh of a specific face with m vertices. The pose parameters consist of the scale factor f, the 3D rotation matrix $R \in \mathbb{R}^{3 \times 3}$ and the 3D translation $t \in \mathbb{R}^3$. $S \in \mathbb{R}^{3 \times m}$ represents the pose-invariant face shape, which is disentangled into the mean shape template $\overline{S} \in \mathbb{R}^{3 \times m}$ from [26] and the deformation $D \in \mathbb{R}^{3 \times m}$ between the actual shape S and the mean shape $\overline{S}$.

3.3. Dual face regression network

Following the approach proposed in [25], we adopt a joint regression strategy to simultaneously estimate the pose-invariant face S and the pose-dependent face P. Then, we employ a self-alignment module denoted by ϕ to estimate face pose from the estimated faces:

$$\phi(P, S) = f, R, t. \tag{6}$$

Based on the estimated face pose $\phi(P, S)$ and pose-invariant face S, we could reconstruct the final shape G via transformation defined in Eq. (4).

The joint regression of the pose-invariant face S and pose-dependent face P is complimentary. For one thing, it helps to separate the effects of non-rigid facial changes due to expression from those resulting from rigid facial changes due to head pose. For another, the learning of pose-dependent face is easy to over-fit to pose and under-fit to facial expressions as the pose variations bring much greater point-to-point distances than the expression variations, resulting in poor reconstruction of expression in various views. By introducing the pose-invariant face S, which remains unchanged with the pose, the network can focus on modeling facial expressions. Meanwhile, as S is disentangled into mean face template $\overline{S}$ and the deformation D, only the zero-centered D is required to be predicted, thereby reducing the complexity of the fitting process.

To facilitate the self-alignment process, the face geometry G, pose-invariant face S, mean face $\overline{S}$, deformation D, and pose-dependent face P are transformed into UV space [27] as UV maps. Since pixels with the same coordinates in the UV maps correspond to the same semantic region, the self-alignment module estimates the similarity transformation matrices using two sets of landmarks extracted from P and S, respectively, based on the same UV coordinate set.

The dual face regression network is pretrained in a supervised manner on the 300W-LP dataset [28]. After the pretraining, we select the branch of the network that generates the pose-invariant face and employ it as our pose normalizer. This component takes a facial image as input and generates the corresponding pose-invariant face.

4. Experimental Setup

4.1. Dataset

The experiments are carried out on LRS3 [29] and MEAD [30] dataset. LRS3 is a large-scale in-the-wild dataset that includes videos obtained from the TED YouTube channel. MEAD contains talking-face videos which are simultaneously recorded at seven different views in a strictly-controlled environment.

In our experiments, we simulate two-speaker mixtures from these two datasets, respectively. The target speech is mixed with a random interference speech at a random signal-to-noise ratio (SNR) in the range from -10 dB to 10 dB. The audio sampling rate is 16 kHz and the face track video of the target speaker is provided at 25 frames-per-second (FPS). For LRS3-2mix, there are 20,000 and 5,000 speech mixture utterances for training and validation simulated from the trainval set, and 3,000 speech mixtures for testing from 168 unseen and unheard speakers during training. For MEAD-2mix, we only select videos with neutral emotion intensity captured from the frontal view for training and validation. The test set contains videos from seven different views, as shown in Figure 2-(c). In total, it consists of 10,000, 1,000 speech mixture utterances for training and validation, and 1,000 for each individual view for testing.

4.2. Implementation details

On LRS3-2mix, we train the whole neural network for 50 epochs with the initial learning rate set to $1e^{-3}$. On MEAD-2mix, we load the weights of the pretrained model on LRS3 to get a good starting point and fine-tune the neural network for 20 epochs with the initial learning rate set to $1e^{-4}$. The learning rate is halved if the accuracy on the validation set does not improve for 3 consecutive epochs. The training process would stop if the accuracy does not improve for 6 consecutive epochs. Adam [31] is used as the optimizer. Gradient clipping with a maximum L2-norm of 5 is applied during training.

5. Results

5.1. Pose-invariant evaluation on MEAD

To analyze the effectiveness of pose-invariant faces in AVSE, we evaluate several variants of PIAVE as shown in Table1. One major impact of pose variations on visually-assisted speech processing systems is decreased performance in mismatched train/test conditions, i.e., the neural network is trained and tested on different poses [32]. Therefore, we conduct evaluations on the MEAD dataset, which includes videos captured from multiple views simultaneously, as illustrated in Figure 2-(c), to assess the performance of different systems under mismatched conditions. Specifically, we train the system using only front-view videos and test it on seven different views.

In Table 1, we use the signal-to-distortion ratio (SDR) to measure the signal quality of the extracted speech under multiple views. For pose-invariant evaluation, we used the average SDR across all seven views (Avg(7)) to measure the overall signal quality and Avg(6) to measure the signal quality under mismatched train/test conditions. Our results indicate that the performance of the model degrades as the view distance from the front one increases, demonstrating the impact of pose variations. PIAVE (w/o PF) performs the worst among all variants, indicating the severe degradation of performance without the pose-invariant faces under pose variations. Comparison between PIAVE (Mask Upper) and PIAVE (Mask Lip) reveals that the movement of lip region have a more significant impact than the upper region of pose-invariant faces. The integration of both components, i.e., the entire face region, yields the best performance in PIAVE, with a 45% average improvement in performance over PIAVE (w/o PF). In conclusion, introducing pose-invariant faces improves the system's performance and makes it more robust against mismatched conditions under pose variations, bringing it one step closer to pose-invariant AVSE.

5.2. Ablation study of PIAVE on LRS3

In Table 2, we present an ablation study on LRS3, a large-scale in-the-wild dataset, to elaborate on the significance of different video streams. Besides SI-SDR and SDR to measure the signal quality, we use the perceptual evaluation of speech quality (PESQ) [33] and the short term objective intelligibility (STOI) [34] to evaluate the perceptual quality and intelligibility of the extracted speech. The higher the better for these metrics.

To illustrate the effect of incorporating pose-invariant faces, PIAVE brings about 0.52dB SI-SDR and 0.143 PESQ improvement when compared to PIAVE (w/o PF). When the lip region of pose-invariant faces is masked in PIAVE (Mask Lip), the resulting performance is even worse than that of PIAVE (w/o PF) in terms of SI-SDR and SDR. This result highlights that primarily the lip movements of pose-invariant faces contribute to the performance improvement. Additionally, PIAVE (Mask Upper) outperforms PIAVE (w/o PF), but there is still a gap between PIAVE (Mask Upper) and PIAVE, emphasizing the need for the utilization of the entire face region. Specifically, the upper region of the talking face offers a coarse-grained synchronization cue, while the lip region provides fine-grained viseme-phoneme mapping. We note that when the amount of training data is limited, the benefits of the upper region of the talking face can be larger, as depicted in Table 1. In summary, our results show that for the dataset with various pose variations, providing the model with stable pose-invariant visual input helps machine lis-

Table 1: *Pose-invariant evaluation on MEAD dataset. Performance is reported with SDR(dB). Different views correspond to Figure 2-(c). PF refers to input with pose-invariant faces, and the masking regions in PIAVE (Mask Lip) and PIAVE (Mask Upper) is illustrated in Figure 2-(b). Avg(7) refers to the average SDR value of all 7 different views. Avg(6) refers to the average SDR value of 6 different views except for the front view.*

Model	Front	Top	Down	Left_30	Left_60	Right_30	Right_60	Avg(7)	Avg(6)
PIAVE (w/o PF)	9.712	5.641	5.698	4.888	1.875	7.226	4.532	5.653	4.977
PIAVE (Mask Lip)	10.277	7.078	5.301	5.328	5.804	5.277	5.107	6.310	5.649
PIAVE (Mask Upper)	9.974	6.615	6.718	8.102	5.951	8.170	**6.142**	7.382	6.950
PIAVE	**11.773**	**8.923**	**8.514**	**8.583**	**6.118**	**8.387**	4.935	**8.176**	**7.577**

Table 2: *Ablation study on LRS3 dataset. PF refers to input with pose-invariant faces, and the masking regions in PIAVE (Mask Lip) and PIAVE (Mask Upper) is illustrated in Figure 2-(b).*

Sys.	SI-SDR	SDR	PESQ	STOI
PIAVE (w/o PF)	14.734	15.064	2.442	0.937
PIAVE (Mask Lip)	14.573	14.906	2.461	0.937
PIAVE (Mask Upper)	14.970	15.279	2.544	0.942
PIAVE	**15.255**	**15.569**	**2.585**	**0.944**

Table 3: *Comparison of our PIAVE model with the baseline models on LRS3 dataset.*

Model	SI-SDR	SDR	PESQ	STOI
Visualvoice [8]	9.603	10.112	2.089	0.897
TDSE [7]	13.253	13.778	2.351	0.912
PIAVE (Ours)	15.255	15.569	2.585	0.944

tening. Furthermore, integrating pose-invariant faces and original face tracks results in an improved observation of the talking face from multiple views, including lip movements and facial expressions from the upper region of the face, in which lip movements play a major role.

5.3. PIAVE vs baseline on LRS3

In Table 3, a comparison of the results of our PIAVE model with two recent AVSE models, VisualVoice [8] and TDSE [7], is presented. VisualVoice is a complex spectral mapping approach that is used for speech separation, and TDSE is a time-domain speaker extraction system. We reproduce the aforementioned two baseline models and evaluated them on the same dataset used for training and testing PIAVE, to ensure a fair comparison. The results demonstrate that PIAVE outperforms the current SOTA models in terms of speech quality and intelligibility.

6. Discussion

Head pose variations affect the identification of the auditory component of audio-visual speech stimuli [35], and how to reduce their impact has been the focus of previous research. This includes attempts to develop pose-invariant lipreading through the learning of multiple classifiers for each specific poses [36] or with a mapping function to transform features to a specific view, e.g. $30°$, $45°$ and $60°$ to the frontal view [37] or to the $30°$ view [38]. However, these approaches have their limitations in terms of being time-consuming with the increase of pose numbers and constrained by specific pose orientations. Recent research in lipreading has explored the efficacy of using multiple views of lips for improved performance [32]. While this approach has shown to be beneficial, it is limited by the need

for multiple cameras. Moreover, recent research has investigated the impact of head movements on audio-visual speech enhancement and demonstrates the effectiveness of using face frontalization to remove head movements [39]. However, this approach is only validated on frontal views.

We present the first attempt to address the pose variation problem in audio-visual speaker extraction with a pose-invariant view. Unlike previous approaches to this problem, we generate a corresponding pose-invariant view for any visual input with only one camera to enable stable visual input and multi-view observation. We believe this work can provide inspiration for other visually-assisted speech processing algorithms.

There are also many open future directions. In the current stage, the pose-invariant face generated by the pose normalizer lacks facial texture, which typically contains identity information such as gender. This limitation may offset the advantages of pose normalization as a cue for the model to accurately identify the target speaker. Furthermore, potential areas for future research also include exploring more effective techniques for feature fusion between two video streams, as well as between audio and visual modality.

7. Conclusion

In this paper, we have proposed a Pose-invariant Audio-visual Speaker Extraction Network (PIAVE) to address the pose variation problem, which is largely unexplored in AVSE. Specifically, we generate the pose-invariant view from each original pose orientation, which enables the model to receive a consistent frontal view of the talker regardless of his/her head pose. We validate the effectiveness of PIAVE on two datasets: the large-scale in-the-wild LRS3 dataset and the multi-view talking face MEAD dataset. Our experimental results demonstrate that the incorporation of pose-invariant faces results in a more robust model capable of handling variations in the head pose. It also improves the overall performance, by enabling stable input of the visual modality, as well as multi-view observation of talking faces. In summary, PIAVE provides a practical solution for the pose variation problem and is a step forward in modeling the cocktail party effect in uncontrollable circumstances.

8. Acknowledgments

9. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voice-filter: Targeted voice separation by speaker-conditioned spectrogram masking," *Interspeech*, pp. 2728–2732, 2019.

[5] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1370–1384, 2020.

[6] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," *Interspeech*, pp. 1406–1410, 2020.

[7] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *ASRU*. IEEE, 2019, pp. 667–673.

[8] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *CVPR*. IEEE, 2021, pp. 15 490–15 500.

[9] Q. Liu, Y. Huang, Y. Hao, J. Xu, and B. Xu, "Limuse: Lightweight multi-modal speaker extraction," in *SLT*. IEEE, 2023, pp. 488–495.

[10] J. Li, M. Ge, Z. Pan, L. Wang, and J. Dang, "VCSE: Time-Domain Visual-Contextual Speaker Extraction Network," in *Interspeech*, 2022, pp. 906–910.

[11] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information." in *Interspeech*, 2019, pp. 4290–4294.

[12] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-spex: Localized target speaker extraction," in *ICASSP*. IEEE, 2022, pp. 7287–7291.

[13] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[14] Z. Pan, M. Ge, and H. Li, "Usev: Universal speaker extraction with visual cue," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 3032–3045, 2022.

[15] M. Baldassare, "Human spatial behavior," *Annual Review of Sociology*, vol. 4, no. 1, pp. 29–56, 1978.

[16] S. Cheng, P. Ma, G. Tzimiropoulos, S. Petridis, A. Bulat, J. Shen, and M. Pantic, "Towards pose-invariant lip-reading," in *ICASSP*. IEEE, 2020, pp. 4357–4361.

[17] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

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

[19] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[20] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP*. IEEE, 2019, pp. 626–630.

[21] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *CVPR*, 2015, pp. 787–796.

[22] M. He, J. Zhang, S. Shan, M. Kan, and X. Chen, "Deformable face net: Learning pose invariant feature with pose aware feature alignment for face recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

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

[24] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[25] Z. Ruan, C. Zou, L. Wu, G. Wu, and L. Wang, "Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction," *IEEE Transactions on Image Processing*, vol. 30, pp. 5793–5806, 2021.

[26] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.

[27] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *ECCV*, 2018, pp. 534–551.

[28] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016, pp. 146–155.

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

[30] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*. Springer, 2020, pp. 700–717.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[32] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multi-view lipreading," *British Machine Vision Association and Society for Pattern Recognition*, 2017.

[33] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2. IEEE, 2001, pp. 749–752.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*. IEEE, 2010, pp. 4214–4217.

[35] T. R. Jordan and K. Bevan, "Seeing and hearing rotated faces: influences of facial orientation on visual and audiovisual speech recognition." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 23, no. 2, p. 388, 1997.

[36] N. Hesse, T. Gehrig, H. Gao, and H. K. Ekenel, "Multi-view facial expression recognition using local appearance features," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 3533–3536.

[37] V. Estellers and J.-P. Thiran, "Multipose audio-visual speech recognition," in *2011 19th European Signal Processing Conference*. IEEE, 2011, pp. 1065–1069.

[38] Y. Lan, B.-J. Theobald, and R. Harvey, "View independent computer lip-reading," in *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 2012, pp. 432–437.

[39] Z. Kang, M. Sadeghi, R. Horaud, X. Alameda-Pineda, J. Donley, and A. Kumar, "The impact of removing head movements on audio-visual speech enhancement," in *ICASSP*. IEEE, 2022, pp. 7302–7306.