Speech Synthesis Model Based on Face Landmarks

Chenji Jin
Hangzhou Dianzi University
epsiotapi@hdu.edu.cn

Yoshimi Suzuki
The university of Yamanashi
ysuzuki@yamanashi.ac.jp

Fei Lin
Hangzhou Dianzi University
linfei@hdu.edu.cn

Abstract

Lip reading recognition aims to predict what people are saying based on the movements of their lips. Most previous works used continuous images to represent lip movements and predict the corresponding textual contents, which does not achieve good performance. In this work, we explore a new approach to synthesizing audio through lip movements by introducing face landmarks for representing the motion features of the face in 3D space and synthesizing the corresponding audio results directly. We propose the FaceLandmarks2Wav model for a preliminary implementation of the above idea. The experimental results confirm that face landmarks can adequately represent facial movement features, and the structure of FaceLandmarks2Wav could synthesize speech results close to natural human voices, only using the face landmarks sequence.

1 Introduction

The relationship between human speech lip movements and pronunciation has been confirmed in many previous studies (Cappelletta and Harte, 2012; Shaikh et al., 2010). Trained professionals can predict what others say by observing their lip shape. People with hearing impairment use a similar method to understand what others say when communicating. Visual speech recognition uses the relationship between lip movements and pronunciation to predict what a speaker says by capturing videos of their lip position. Related research has many practical applications, such as assisting people with acquired aphasia to communicate with others.

Research related to computer lip recognition generally extracts lip gesture features from continuous images of videos (Ma et al., 2022; Huang et al., 2022; Wang et al., 2022). Early studies mainly used image transformation methods to reduce the dimensionality of feature vectors. (Min and Zuo, 2011) performed lip visual feature extraction based on 3D-DCT and 3D-HMM models, which focused on the primary information of images in the low-frequency band. With the development of research in computer vision, the extraction of lip movement features in images using deep learning networks such as CNN (Iezzoni et al., 2004; Fung and Mak, 2018; NadeemHashmi et al., 2018; Chung and Zisserman, 2016) has also received increasing attention. Noda et al. (2014) used a CNN-based multilayer network to extract feature sequences from lip images and modeled them by GMM-HMM. Garg et al. (2016) used LSTM networks to extract lip movement in the temporal dimension information.

Current research has primarily used continuous images of the speaker's face to illustrate lip movements. However, video is not the most intuitive way to represent lip movements. The video samples contain much redundant information, requiring a large-scale network to locate the speaker's lips and extract movement features accurately. Even then, redundant information can also interfere with model predictions. For example, the model relies on the facial details of the speaker in the video and may not make accurate estimations when encountering an unseen speaker, which is more common when training the model with person-specific video datasets. In addition, the possible facial rotation of people while speaking can cause the camera to not continuously capture the face of the picture, which limits the application scenarios of lip recognition research. To solve the problems in video lip recognition, we introduce face landmarks to represent facial movement states in lip recognition. Face landmarks are a series of coordinate points annotated on the human face, often used to track the positional states of facial features.

In this study, we combine face landmarks in the temporal dimension into a sequence to represent the movement features of the face in three-dimensional space. Extract facial movement features from fa-

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 48–54
November 1–4, 2023. ©2023 Association for Computational Linguistics

cial landmarks sequence by an encoder consisting of multilayer convolutional neural networks and LSTM, and use an autoregressive decoder to synthesize the audio close to the speaker's pronunciation.

2 Methods

We refer to the method used by Shen et al. (2018) in the text-to-speech task. Our model does not directly synthesize the audio waveform from the lip movement sequence. Instead, it predicts the mel spectrogram of the corresponding audio segment and uses a vocoder to convert the mel spectrogram to audio results. Assume that the lip movement sequence are represented as $L = (L_1, L_2, \cdots, L_T)$ and the mel spectrogram are represented as $M = (M_1, M_2, \cdots, M_{T'})$, and the mel spectrogram are represented as follows:

$$S = (S_1, S_2, \cdots, S_T),\ S_i = (s_1, \cdots, s_N) \quad (1)$$

$$L = (L_1, L_2, \cdots, L_{T'}),\ L_i = (P_1, \cdots, P_F) \quad (2)$$

Where T and T' are the frame numbers of the lip movement sequence and the mel spectrogram in the same video clip, F is the number of 3D landmarks in the facial feature part selected in the experiment. N is the number of mel filters.

We can assume that the representation of the target mel spectrogram in the t' frame is highly correlated with the lip movements of the speaker at the exact moment. However, since there are possibilities where different phonemes share the same viseme, to determine the mel spectrogram of the t' frame, the model should also reference the context of the lip movement feature. We model the relationship between the mel spectrogram and lip movements using Eq. (3).

$$M_{t'} = f\left(L_{k\in(t\pm\delta)}, M_{<t'}\right) \quad (3)$$

The encoder refers to context information to extract lip movement features. The decoder uses an autoregressive method to synthesize the corresponding mel spectrogram frame by frame. The model structure is shown in Fig. 1.

2.1 Input/Output Representation

2.1.1 Input Representation

FaceLandmarks2Wav model accepts face landmarks as input. Each landmark contains three-dimensional coordinate information. Therefore, the tensor size of the model input is $T \times F \times 3$, where F is the number of landmarks used in the experiment, and T is the number of time steps of the landmarks sequence, each training sample uses 90 continuous frames of image content.

2.1.2 Output Representation

The target synthesized by our model is the audio content corresponding to the given video segment. FaceLandmarks2Wav does not directly synthesize audio results but predicts the corresponding mel spectrogram. We sample the audio at a sampling rate 16kHz, set the window size to 50ms, shift distance per frame to 12.5ms, and set the number of mel filters to 80. As the model obtained the corresponding mel spectrogram, we use the Griffin-Lim algorithm (Griffin and Lim, 1984) to transform it into the corresponding audio wave.

2.2 Spatio-temporal Face Encoder

Previous studies (Bai et al., 2018; Xu et al., 2019) have demonstrated the effectiveness of CNN for feature extraction in the time domain. Therefore, we stacked convolutional blocks to extract movement features from face landmarks input. The input dimension of the encoder is $T \times F \times 3$. Unlike processing images, the number of face landmarks F corresponds to different convolution channels, which helps the convolution kernel respond to all facial landmarks.

We set multiple convolution blocks in the encoder, and each convolution block increases the number of channels used to represent facial features. The number of channels of each convolution block is set according to the face landmarks used in the experiment. Residual connections and batch normalization are used between CNN blocks. The last layer of convolutional blocks will sample the three-dimensional coordinate information into one dimension, and the encoder will permanently preserve the time dimension. The convolutional network's final output size is $T \times F'$, and F' is the number of features modeled by the encoder for a single time step.

The encoder uses a bidirectional LSTM network (Hochreiter and Schmidhuber, 1997) to extract short-term contextual features. This method allows the feature modeling of face landmarks to contain more contextual information after the convolutional network.

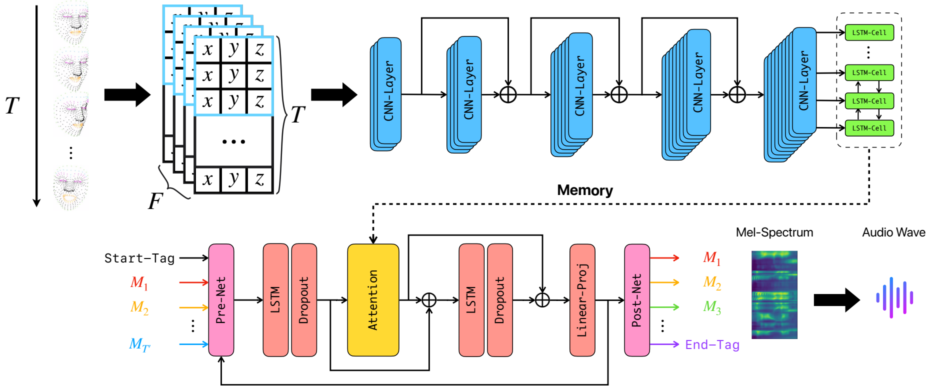


Figure 1: FaceLandmarks2Wav model structure. The encoder uses a 2D convolutional network to extract high-level lip movement features from landmarks. The decoder predicts the mel spectrogram corresponding to the audio result autoregressively.

2.3 Attention-based Speech Decoder

To synthesize smoother and natural speech results, our model refers to the method used in Tacotron2 by Shen et al. (2018). The Tacotron2 is a model for synthesizing audio from text, which uses a sequence-to-sequence network with an attention mechanism to process the text features extracted by the encoder and synthesize a mel spectrogram close to the natural human voice. We use a Tacotron2-like decoder to autoregressively synthesize the mel spectrogram frame by frame from the facial movement features encoded by the encoder. When the decoder synthesizes the mel spectrograms output of the T_k time step, it will refer to the decoder output of T_{k-1} time step and calculates the attention together with the lip movement high-level features extracted by the encoder.

The attention network contains a special location layer (Chorowski et al., 2015), which accepts the accumulated attention weights from previous time steps as an additional condition, which can help the attention network to calculate the attention weights forward and prevent the decoder from falling into repeated patterns. Such a decoder structure contributes to more natural audio results for model synthesis.

2.4 Loss Function

The optimization goal during model training is minimizing the hybrid loss between the synthetic mel spectrogram and the ground truth for end-to-end model training. The hybrid loss function is shown in Eq. 4, and α is the weight used to adjust the loss function and is set to 0.5 in the experiment.

$$\mathcal{L}_{all} = \alpha \cdot \mathcal{L}_1 + \mathcal{L}_{MSE} \tag{4}$$

3 Benchmark Datasets and Training Details

3.1 Datasets

Accurate 3D face landmarks can be annotated on faces using special devices such as the True Depth camera on the iPhone. However, since face landmark is a novel way to describe lip movements, there is no previous research on lip recognition using similar methods. To compare with those studies using video data, we use the face landmarks extractor to extract the face landmarks from the existing video dataset. Build the face landmarks dataset based on the video dataset. We chose the Lip2Wav dataset from Prajwal et al. (2020) as the source of lip recognition video data, which collects about 120 hours of video data from Youtube, including facial images of different speakers when they spoke and divided them into different sub-datasets according to the different speakers in the video. It is very suitable for the model to learn the lip synthesis style of a specific speaker.

We chose the Media Pipe Face Mesh model (MPFM) proposed by Grishchenko et al. (2020) as the face landmarks extractor. The MPFM model could provide 478 3D landmark coordinates of the whole face range. This model also optimizes the face landmarks labeling of continuous images and reduces the jitter of landmarks between frames. Those features make the MPFM model more suitable for labeling face landmarks on video data.

After getting the face landmarks on the video data, we normalize the sequence of time-series face landmarks using Eq.(5). This function will make the coordinates of the face landmarks have appropriate sparsity.

$$\begin{aligned} \boldsymbol{c} &\leftarrow \frac{1}{3n}\sum_{i=1}^{n}\sum_{j=1}^{3}\boldsymbol{p}_{ij} \\ m &\leftarrow \max\left(||\boldsymbol{p}_{ij}-\boldsymbol{c}||\right) \\ \boldsymbol{p}'_{ij} &\leftarrow \frac{1}{m}(\boldsymbol{p}_{ij}-\boldsymbol{c}) \end{aligned} \tag{5}$$

Since the video in the Lip2Wav dataset is captured by fixed camera position, some of the video clips cannot contain the range of the human face, and this part of the samples can not be used in training. To ensure that the encoder can better model the lip-movement contextual features, we set the video window to 90 frames, and this means that only face landmarks extractors can recognize 90 consecutive face frames will be used as samples for model training. The video lip-synthesis model used for comparison was also trained using the same data range, and the number of samples contained in each sub-dataset is shown in Table 1.

Table 1: The number of samples in each sub-dataset when using different face landmarks extractors.

sub-datasets	Media Pipe FaceMesh	Face Alignment
chem	753,834	572,198
chess	787,963	547,220
eh	819,065	739,483

3.2 Details of Training

We use the pre-trained face landmarks extraction model to recognize and mark faces in the video data frame by frame. When using Media Pipe Face Mesh, the video mode will be turned on to reduce the jitter of landmarks. Our model sets a multi-layer convolutional network in the encoder, which finally represents the movement features of each frame as a high-dimensional feature vector as the hidden dimension of the encoder. Taking the experiment with 80 lip landmarks as an example, we sequentially set the number of channels of the convolutional block in the encoder to 120, 240, and 320, and the final hidden dimension is set to 384.

The batch size during model training is set to 32. The learning rate will increase linearly to 0.001 at the beginning of training and gradually decay in subsequent iterations. We used Adam (Kingma and Ba, 2014) as the optimizer and trained the model with about 600,000 iterations. The choice of these parameters was derived from the good results obtained during the experiments.

3.3 Evaluation Metrics

We measure whether the audio synthesized by the model is close to the natural human voice regarding intelligibility and audio quality. We will use the following three metrics to compare our model with previous studies: Short-Term Objective Intelligibility (STOI) (Taal et al., 2010), Extended Short-Term Objective Intelligibility (ESTOI) (Jensen and Taal, 2016), and Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001). In addition to the objective audio quality, we will compare the model's resource consumption and convergence time during the training process and demonstrate the comprehensive advantages of the unique solution of using face landmarks to represent lip movements from many aspects.

4 Results and Disscussion

This section presents a comparative analysis to evaluate the performance differences between FaceLandmarks2Wav and previous deep-learning models that use videos for lip-speech recognition. We choose the Lip2Wav as the baseline model (Prajwal et al., 2020), a lip-speech model that can synthesize audio results close to natural human voice from facial videos. We trained different models on the same dataset using similar settings and compared the differences in the number of parameters, Multiply-Accumulate operations (MACs), and convergence time of the models during training. The results are shown in Table 2. We used a smaller batch size when training Lip2Wav due to the limited video memory capacity of the graphics card used, and even so, the FaceLandmarks2Wav also has more advantages in terms of convergence time and other metrics.

Table 2: Comparison of model details and training time.

Models	Lip2Wav	Ours
Batch Size	16	32
Parameters	39.8 M	31.9 M
MACs	709.3 G	178.4 G
Single Iteration	1.4 sec	0.4 sec
Convergence	$\sim$ 140 hours	$\sim$ 50 hours

During training, we record the attention image generated by the FaceLandmarks2Wav. The horizontal and vertical axes of the image represent the time steps of the decoder and the encoder, respectively. The values represent the degree of attention paid to the encoder's specific time step when the decoder's attention module produces the corresponding time step results. Figure 2 shows how the attentional alignment of the model changes during the training process. The attention image gradually forms a diagonal image as the training progresses, meaning that the decoder refers to the lip features extracted by the encoder at nearby moments when synthesizing the audio results, consistent with the assumption of Eq 3. The attention image at the end of training is shown in Figure 2f, implying that the model could already learn the high-level features of facial motion from the input face landmarks sequence and synthesize the corresponding audio based on these features.

After the training, we compared the synthetic audio quality and intelligibility scores of the two models on the validation set, and the results are shown in Table 3. Compared with Lip2Wav, Our FaceLandmarks2Wav has a significant advantage in STOI and PESQ scores, and the ESTOI scores of Lip2Wav are relatively better. While synthesizing high-quality audio, FaceLandmarks2Wav has shorter training time, inference time, and smaller model sizes than Lip2Wav. Therefore, in scenarios sensitive to video memory usage and requiring high real-time performance, Our approach of synthesizing audio using face landmarks has more advantages.

Discussion. The above experiments prove that using face landmarks can represent the attributes of facial movement well. The experimental results show that using the FaceLandmarks2Wav model can synthesize natural and smooth speech results, and the synthesized audio is not inferior to the model using video data as input in terms of intelligibility and quality. Our proposed model structure can converge faster during the training process, and the requirements for the training environment are further reduced. The small model size allows it to be trained in environments with limited hardware, such as wearable devices. We also implemented corresponding ablation studies to compare the effect of face landmarks extracted differently on model performance.

Table 3: Performance comparison between our model and previous lip speech synthesis studies. The column of total result shows the arithmetic mean of the results of different sub-datasets.

Sub-dataset	**Metrics**	**Models**	
		Lip2Wav	**Ours**
chem	STOI	0.414	**0.478**
	ESTOI	**0.212**	0.193
	PESQ	1.130	**1.149**
chess	STOI	0.168	**0.217**
	ESTOI	**0.101**	0.073
	PESQ	1.143	**1.151**
eh	STOI	0.256	**0.367**
	ESTOI	**0.012**	0.009
	PESQ	1.302	**1.318**
total result	STOI	0.279	**0.354**
	ESTOI	**0.105**	0.092
	PESQ	1.192	**1.201**

5 Ablation Studies

As an initial study of using face landmarks to represent facial movement features, the effect of different ranges of face landmarks on the ability to represent facial motion features is one of our primary concerns, and we designed the corresponding ablation studies to explore this question.

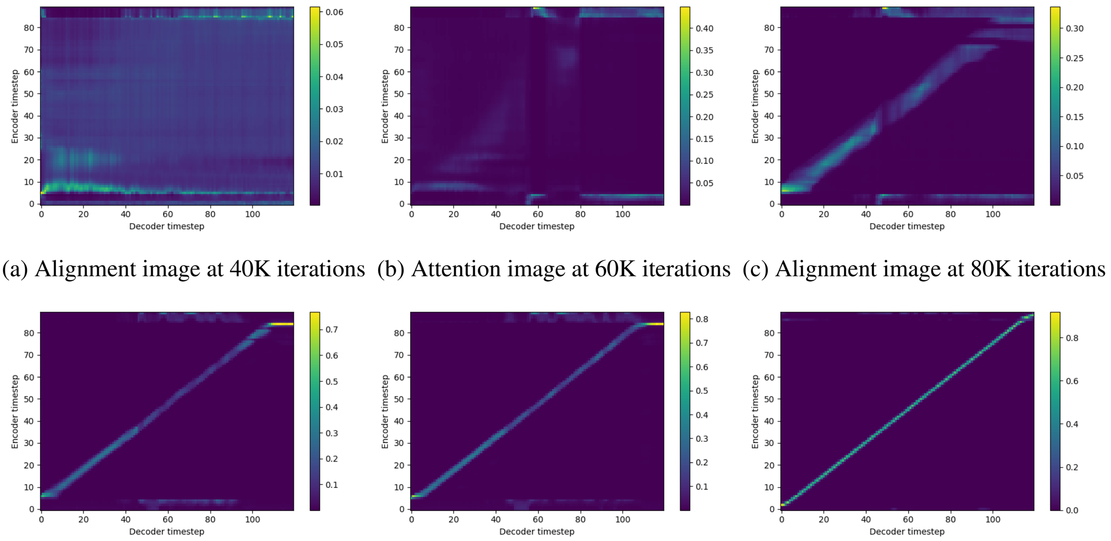
We also try to use the Face Alignment Network (FAN) proposed by Bulat and Tzimiropoulos (2017) as an alternative face landmarks extractor, which can annotate 68 2D-landmarks in the whole face range and provide a way to estimate the depth information.

Table 4: Train the model with different ranges of landmarks. Training is performed on the chem sub-dataset. WF stands for "Whole Face".

Extractor	**MPFM**	**MPFM**	**FAN**
Contents	WF	Lips	WF
Landmarks	478	80	68
Channel	600, 680, 720	120, 240, 320	
Embedding	768	384	384
STOI	0.348	**0.478**	0.372
ESTOI	0.109	**0.193**	0.103
PESQ	1.048	**1.149**	1.034

The results of the ablation studies are shown in Table 4, which also shows the model embedding parameters for different settings of the number of landmarks. Surprisingly, the training results using only the lip landmarks are superior to those

Figure 2: Attention alignment images for FaceLandmarks2Wav



(a) Alignment image at 40K iterations (b) Attention image at 60K iterations (c) Alignment image at 80K iterations

(d) Attention image at 100K iterations (e) Attention image at 120K iterations (f) Alignment image at end of training

using the whole-face range landmarks. This phenomenon is because the target audio synthesized by the model is most closely associated with the lip movements, and the input contains less additional information, making it more advantageous to train directly on the lip content. The experimental results do not show significant differences for different extractors using different numbers of landmarks to represent the full-face range of motion features.

On the other hand, when using the same face landmarks extractor, the performance of 3D landmarks is significantly better than that of 2D landmarks, indicating that even the depth information estimated by the extractor still provides more effective facial motion information to the encoder. This phenomenon is important for our future work, which means that the ability of face landmarks for facial motion representation could be further improved if landmark annotation is performed directly on real faces using a custom device.

6 Conclusion and Future Work

In this study, we initially explored the possibility of an innovative approach to characterize facial motion using face landmarks. We proposed FaceLandmarks2Wav, a model that synthesizes corresponding lip reading audio based on face landmarks and compared it with Lip2Wav, the lip reading model that uses video data to synthesize audio results. Experimental results show that our proposed model structure can synthesize relatively natural and smooth audio structures and be trained in a lower hardware environment. We also performed ablation studies, showing that audio results synthesized only using the lip range are even better than those using the whole face range. In future work, we hope to directly obtain the depth information of lip movement through 3D camera equipment, and more accurate face landmarks information will help further improve the model's performance.

References

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030.

Luca Cappelletta and Naomi Harte. 2012. Phoneme-to-viseme mapping for visual speech recognition. In *ICPRAM (2)*, pages 322–329. Citeseer.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28.

Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer.

Ivan Fung and Brian Mak. 2018. End-to-end low-resource lip-reading with maxout cnn and lstm. In

2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2511–2515. IEEE.

Amit Garg, Jonathan Noyola, and Sameep Bagadia. 2016. Lip reading using cnn and lstm. *Technical report, Stanford University, CS231 n project report.*

Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243.

Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. 2020. Attention mesh: High-fidelity face mesh prediction in real-time. *arXiv preprint arXiv:2006.10962*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hongyang Huang, Chai Song, Jin Ting, Taoling Tian, Chen Hong, Zhang Di, and Danni Gao. 2022. A novel machine lip reading model. *Procedia Computer Science*, 199:1432–1437.

Lisa I Iezzoni, Bonnie L O'Day, Mary Killeen, and Heather Harker. 2004. Communicating about health care: observations from persons who are deaf or hard of hearing. *Annals of internal medicine*, 140(5):356–362.

Jesper Jensen and Cees H Taal. 2016. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2022. Visual speech recognition for multiple languages in the wild. *arXiv preprint arXiv:2202.13084*.

Kim Yong Min and Li Hong Zuo. 2011. A lip reading method based on 3-d dct and 3-d hmm. In *Proceedings of 2011 International Conference on Electronics and Optoelectronics*, volume 1, pages V1–115. IEEE.

Saquib NadeemHashmi, Harsh Gupta, Dhruv Mittal, Kaushtubh Kumar, Aparajita Nanda, and Sarishty Gupta. 2018. A lip reading model using cnn with batch normalization. In *2018 eleventh international conference on contemporary computing (IC3)*, pages 1–6. IEEE.

Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. 2014. Lipreading using convolutional neural network. In *fifteenth annual conference of the international speech communication association*.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Ayaz A Shaikh, Dinesh K Kumar, Wai C Yau, MZ Che Azemin, and Jayavardhana Gubbi. 2010. Lip reading using optical flow and support vector machines. In *2010 3Rd international congress on image and signal processing*, volume 1, pages 327–330. IEEE.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE.

Huijuan Wang, Gangqiang Pu, and Tingyu Chen. 2022. A lip reading method based on 3d convolutional vision transformer. *IEEE Access*, 10:77205–77212.

Kai Xu, Longyin Wen, Guorong Li, Liefeng Bo, and Qingming Huang. 2019. Spatiotemporal cnn for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1379–1388.