applied sciences

MDPI

Article

Lip2Speech: Lightweight Multi-Speaker Speech Reconstruction with Gabor Features

Zhongping Dong [1], Yan Xu [1], Andrew Abel [2,*] and Dong Wang [3]

[1] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China; sally.dongzp@gmail.com (Z.D.); yan.xu@xjtlu.edu.cn (Y.X.)
[2] Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XQ, Scotland, UK
[3] Center for Speech and Language Technologies (CSLT), BNRist at Tsinghua University, Beijing 100084, China; wangdong99@mails.tsinghua.edu.cn
* Correspondence: andrew.abel@strath.ac.uk

Abstract: In environments characterised by noise or the absence of audio signals, visual cues, notably facial and lip movements, serve as valuable substitutes for missing or corrupted speech signals. In these scenarios, speech reconstruction can potentially generate speech from visual data. Recent advancements in this domain have predominantly relied on end-to-end deep learning models, like Convolutional Neural Networks (CNN) or Generative Adversarial Networks (GAN). However, these models are encumbered by their intricate and opaque architectures, coupled with their lack of speaker independence. Consequently, achieving multi-speaker speech reconstruction without supplementary information is challenging. This research introduces an innovative Gabor-based speech reconstruction system tailored for lightweight and efficient multi-speaker speech restoration. Using our Gabor feature extraction technique, we propose two novel models: GaborCNN2Speech and GaborFea2Speech. These models employ a rapid Gabor feature extraction method to derive low-dimensional mouth region features, encompassing filtered Gabor mouth images and low-dimensional Gabor features as visual inputs. An encoded spectrogram serves as the audio target, and a Long Short-Term Memory (LSTM)-based model is harnessed to generate coherent speech output. Through comprehensive experiments conducted on the GRID corpus, our proposed Gabor-based models have showcased superior performance in sentence and vocabulary reconstruction when compared to traditional end-to-end CNN models. These models stand out for their lightweight design and rapid processing capabilities. Notably, the GaborFea2Speech model presented in this study achieves robust multi-speaker speech reconstruction without necessitating supplementary information, thereby marking a significant milestone in the field of speech reconstruction.

Keywords: speech reconstruction; lipreading; gabor features; lip features; speech synthesis; image processing; machine learning

Citation: Dong, Z.; Xu, Y.; Abel, A.; Wang, D. Lip2Speech: Lightweight Multi-Speaker Speech Reconstruction with Gabor Features. *Appl. Sci.* **2024**, *14*, 798. https://doi.org/10.3390/app14020798

Academic Editor: Douglas O'Shaughnessy

Received: 17 October 2023
Revised: 4 December 2023
Accepted: 7 December 2023
Published: 17 January 2024

1. Introduction

Speech reconstruction refers to the generation of audio signals from silent lip movement videos [1]. As human beings, speech is our primary mode of communication, but in scenarios such as noisy environments or surveillance videos, speech may be disrupted, masked, or even absent altogether [2]. In these cases, visual information such as facial or lip movements may compensate for missing or disrupted speech signals, thereby enhancing speech perception [3]. The relationship between audio and visual modalities has been demonstrated by the McGurk effect [4]. The focus of this paper, estimating speech from visual information, is known as speech reconstruction [1].

Speech reconstruction differs from speech recognition in that it aims to be a language-independent visual-to-audio task rather than visual and/or audio-to-text [2]. This makes it suitable for applications such as enabling effective communication in noisy environments (e.g., factories or public transportation [5]) or providing artificial speech for patients who

have undergone a laryngectomy. Moreover, speech reconstruction has numerous potential applications as part of a larger communication system, including enhancing speech recognition by filtering out background noise [6,7], improving the quality of speech in assistive hearing devices [8], and generating speech for videos with low-quality or missing sound [9].

However, estimating speech from visual information is challenging due to the incomplete nature of the data obtained from a frontal or side camera. The captured visual information cannot include critical components such as the excitation signal and the majority of tongue movements [10]. Previous work by the authors investigated speech recognition with lipreading using a Chinese language dataset, which included tonal information and found that while Chinese pinyin words could be successfully recognised by lip movements, this was not possible for tonal information only, due to the lack of associated mouth movements [11]. Without access to vocal cords and internal tongue movements, it is challenging to synthesise speech that is intelligible and natural-sounding [10]. Therefore, developing accurate and effective speech reconstruction techniques is an active area of research.

In this paper, a novel Gabor-based speech reconstruction system for lightweight multi-speaker speech reconstruction is proposed. The system utilises Gabor filtering to remove irrelevant facial information and extract key lip features. The extracted features are then fed into an LSTM to model the time dependencies and then a fully connected layer to generate an auditory spectrogram. Compared to end-to-end Convolutional deep neural networks, Gabor feature extraction is a direct and efficient visual extraction process, thus reducing model structure complexity. Using the Gabor features as visual input instead of raw facial pixels significantly reduces the size of the input data and the extracted features can be visualised, making the speech reconstruction system more transparent. The results show that the proposed Gabor-based model achieves an average spectrogram accuracy of 74% in single-speaker scenarios and demonstrates excellent vocabulary reconstruction ability with an average accuracy of 81%. Moreover, the experiments also demonstrate that the proposed GaborFea2Speech model can maintain robust speech reconstruction capability in multi-speaker scenarios, with 72%, 68%, 71%, 65%, and 61% accuracy for different multi-speaker models.

This paper makes several contributions. Firstly, a detailed review of state-of-the-art research is conducted within the visual speech reconstruction domain. Secondly, using our proposed Gabor-based feature extraction method, two advanced lightweight speech reconstruction models are introduced: GaborCNN2Speech and GaborFea2Speech. Thirdly, comprehensive testing across single-speaker, multi-speaker, and vocabulary contexts demonstrates that our proposed GaborCNN2Speech model offers enhanced performance compared to traditional end-to-end Convolutional Neural Network (CNN) models, presenting a combination of superior results, lightweight design, and rapid processing. Moreover, our innovative GaborFea2Speech model capitalizes on three-dimensional lip features, achieving lightweight and rapid system attributes and excelling in single-speaker and vocabulary reconstruction tasks. Most significantly, it achieves multi-speaker speech reconstruction without the need for supplemental information.

The remainder of this paper is presented as follows. Section 2 presents a comprehensive literature review of the speech reconstruction field. Section 3 introduces the Gabor feature extraction technique. Section 4 introduces our novel Gabor-based speech reconstruction models. The experimental methodology is provided in Section 5, with results in Section 6 and discussion in Section 7. Finally, Section 8 concludes the paper and provides future research directions.

2. Literature Review

Initial research into speech reconstruction faced computational limitations [5,12,13]. A full reconstruction was hard to achieve, and so the focus was on low dimensional estimation such as using DCT visual information to estimate filterbank audio features [12–15]. However, increased computational power and more advanced machine-learning models allowed for improved estimation. Early research [12] reconstructed audio from video using

a deep-learning network trained on visual features from the mouth region. Recent work by Akbari et al. [2] utilised an end-to-end deep neural network to learn auditory spectrograms from raw facial pixels. Similarly, Prajwal et al. [16] proposed a sequence-to-sequence architecture that learned individual speaking styles from an unconstrained dataset. However, those approaches are speaker-dependent, which limits their application to a small group of individuals, as each speaker requires a separate model.

Several studies [17–20] have attempted to address the problem of multi-speaker speech reconstruction by adding additional information such as speech or output text to distinguish between speakers. However, if the supplementary information is not present, speakers can not be distinguished. In addition, despite this extra information, these studies [17–19] generally suffer from the problem of low speech-quality reconstruction when reconstructing speech from multiple speakers. Furthermore, the system structure developed in these studies [17–20] is highly complex, which requires a large amount of data and lengthy training, leading to high computational costs. Therefore, there is a need for more lightweight and transparent multi-speaker speech reconstruction models to overcome the limitations of existing models. These models should maintain high speech-quality reconstruction performance without the requirement for additional information, simplifying the system structure.

2.1. Speech Reconstruction Using Constrained Datasets

The first key speech reconstruction system was proposed by Le Cornu and Milner [12], who applied a neural network to estimate a speech signal from a silent video of a speaker's frontal face using a deep-learning network trained on hand-engineered visual features obtained from the mouth region. However, this approach had the limitation of missing certain speech components such as fundamental frequency and aperiodicity. Le Cornu and Milner [5] extended this by using visual features to predict a class label and exploring temporal information using RNNs. Although the intelligibility of the reconstructed speech improved substantially, speech quality was still low. Ephrat and Peleg [1] used a regression-based framework to predict LSP coefficients directly from raw visual data with a CNN and two fully connected layers. They found that no hand-crafted visual features were needed to reconstruct the speaker's voice, using the whole face instead of just the mouth improved performance, and the regression-based method was effective in reconstructing out-of-vocabulary words. However, the signals sounded unnatural because Gaussian white noise was used as excitation to reconstruct the waveform from LPC features.

Subsequent studies focused on improving speech quality and intelligibility. Ephrat and Peleg [21] employed a postprocessing network for transforming a learned mel-scale spectrogram into a linear scale spectrogram. This spectrogram was then used with a Griffin-Lim algorithm [22] to regenerate the time-domain signal. Compared with their early study of Vid2speech [1], their approach demonstrated significant average speech quality improvements. Specifically, during testing on a single GRID speaker, the average PESQ score improved by 38%, from 1.19 to 1.922, as measured by Perceptual Evaluation of Speech Quality (PESQ). Akbari et al. [2] focused on establishing a mapping between the speaker's facial cues and speech-related attributes, leveraging a pre-trained deep Autoencoder network. In comparison to their previous work, their methodology led to a 6% enhancement in average PESQ score and a notable 35% increase in speech intelligibility within the GRID Corpus, encompassing speakers S1, S2, S3, and S4. Similarly, Vougiouskas et al. [23] and Mira et al. [24] utilised Generative Adversarial Networks (GANs) for direct audio waveform synthesis from video frames. During unseen speaker testing, these approaches achieved word error rates of 40.5% with the GRID Corpus and 42.51 with the LRW dataset. Furthermore, Kim et al. [17] integrated a GAN with an attention mechanism, while Yadav et al. [25] employed a stochastic modelling strategy utilising a variational autoencoder. Their methods yielded average PESQ values of 1.961 and 1.932, respectively, with the GRID Corpus. These studies indicate an improvement in PESQ scores over earlier methods, but there are limitations relating to constrained datasets and single speakers. A list of approaches reported in the literature is summarised in Table 1.

Table 1. Review of research into speech reconstruction from silent videos. Here, UD: Unconstrained data, MS: Multi-speaker, MT: Multi-task, MV: Multi-view.

Task	Year	Paper	Input	Output	Model Info	Dataset	Metrics	UD	MS	MT	MV
Reconstructed Speech on Constrained Datasets	2015	Cornu and Milner [12]	2D-DCT/AAM mouth	LPC or mel-filterbank amplitudes	GMM/FFNN	GRID corpus	WER/PESQ	×	×	×	×
	2015	Aihara et al. [13]	2D-DCT	STRAIGHT Spectral	NMF	AV-JP	MOS/PESQ/STOI	×	×	×	×
	2017	Cornu and Milner [5]	AAM mouth	Codebook entries (mel-filterbank amplitudes)	FFNN/RNN	GRID corpus	MSE/ ESTOI/ PESQ	×	×	×	×
	2017	Ephrat et al. [1]	Raw pixels face	Mel-scale and lineae-scale	CNN/FFNN/BiGEU	GRID corpus	Amazon Mechanical Turk (MTurk)	×	×	×	×
	2017	Ephrat and Peleg [21]	Raw pixels face	LSP of LPC	CNN/FFNN	GRID/ TCD-TIMIT	STOI/ ESTOI /PESQ /ViSQOL	×	×	×	×
	2017	Ra et al. [14]	2D-DCT	STRAIGHT Spectral	GMM	M2TINIT	Mel-cepstrum Distortion (MelCD)	×	×	×	×
	2018	Akbari et al. [2]	Raw pixels face	AE features/ spectrogram	CNN/LSTM/FFNN/AE	GRID corpus	PESQ/Corr2D/STMI	×	×	×	×
	2019	Takashima et al. [26]	Raw pixels face	WORLD spectrum	CNN/FFNN	AV-JP	MOS / WER	×	×	×	×
	2019	Vougioukas et al. [23]	Raw pixels face	Raw audio waveform	GAN/CNN/FFNN	GRID corpus	PESQ/WER/ AV Confidence/ AV Offset/STOI/ MCD	×	×	×	×
	2020	Michelsanti et al. [10]	Raw pixels mouth /face	WORLD features	CNN/GRU/FFNN	GRID corpus	PESQ /ESTOI/WER	×	×	×	×
	2021	Joanna et al. [27]	Raw pixels face	Mel spectrograms		GRID/Lip2Wav		×	×	×	×
	2021	Yadav et al. [25]	Raw pixels face	Mel spectrograms	LSTM/AE	GRID corpus	STOI/ ESTOI/ PESQ	×	×	×	×
	2022	Mira et al. [24]	Raw pixels face	Raw audio waveform	CNN/GAN/AE	GRID/LRW/ TCD-TIMIT	PESQ/STOI/MCD/WER	×	×	×	×
	2022	Mira et al. [28]	Raw pixels face	Mel spectrograms	Visual encoder/ SVTS/ Parallel WaveGAN	GRID/LRW,LRS3/ VoxCeleb2	STOI/ESTOI/WER/PESQ	×	×	×	×

Table 1. *Cont.*

Task	Year	Paper	Input	Output	Model Info	Dataset	Metrics	UD	MS	MT	MV
Reconstructed Speech on Unconstrained Datasets	2020	Prajwal et al. [16]	Raw pixels face	AE features/ spectrogram	CNN/LSTM/AE	GRID/TIMIT/ Lip2Wav	CNN/LSTM/AE	✓	×	×	×
	2022	He et al. [29]	Raw pixels face	AE features/ spectrogram	CNN/LSTM/AE	Lip2Wav	STOI /ESTOI/MOS	✓	×	×	×
	2022	Varshney et al. [30]	Raw pixels face	Mel Frequency Cepstral Coefficients (MFCC)	The latent variable model/ transformer	GRID/ Lip2wav	STOI/ ESTOI/ PESQ	✓	×	×	×
	2022	Millerdurai et al. [31]	Raw pixels face	AE features/ spectrogram	CNN/AE/BiLSTM	AVSpeech/LRW/ UTKFce	MOS/STOI/ESTOI/ PESQ /WER	✓	×	×	×
	2022	Wang and Zhao [32]	Raw pixels face	Mel-spectrogram	A spatial-temporal factorized transformer visual encoder	GRID/ Lip2Wav	MOS/PESQ	✓	×	×	×
	2022	Hegde et al. [33]	Raw pixels face	Mel-spectrogram	VAE-GAN	GRID/TCD-TIMIT/ LRW/LRS2	PESQ/ STOI/ SED	✓	×	×	×
Reconstruct Speech from Multi-Speaker	2021	Kim et al. [17]	Raw pixels face	Mel-spectrogram	GAN	GRID/TCD-TIMIT/ LRW	STOI/ESTOI/ PESQ/WER/MOS	×	✓	×	×
	2021	Oneață et al. [18]	Raw pixels mouth	Mel-spectrogram	CNN/ResNet/LSTM	GRID Corpus	STOI / PESQ / MCD/ WER	×	✓	×	×
	2022	Wang et al. [20]	Raw pixels mouth	Mel-spectrogram	VCVTS	GRID/ LRW	PESQ/ STOI/ ESOI / RMSE	×	✓	×	×
	2021	Um et al. [19]	Raw pixels mouth /face	Mel-spectrogram	GAN-based	GRID Corpus	NISQA/ CER/ WER/MOS/	×	✓	×	×
Multi-Task Speech Reconstruction	2019	Qu et al. [34]	Raw pixels face	AE features/ spectrogram	CNN/LSTM/Bi-GRU/ FFNN/AE	GRID Corpus	PESQ/ESTOI/ WER/CER	×	×	✓	×
	2022	Qu et al. [35]	Raw pixels face	AE features/ spectrogram	CNN/LSTM/Bi-GRU/ FFNN/AE	GRID /TCD-TIMIT/ CMLR/LipSound2	PESQ/ESTOI/ WER/CER	×	×	✓	×
	2022	Zeng et al. [36]	Raw pixels face	AE features/ spectrogram	CNN/BiLSTM	GRID/TCD-TIMIT	MOS/ PESQ	×	×	✓	×
	2023	Kim et al. [37]	Raw pixels face	Mel-spectrogram	CNN/ASR model	LRS2/LRS3/LRW	STOI/ ESTOI/ PESQ/ WER	×	×	✓	×
Multi-View Speech Reconstruction	2018	Kumar et al. [38]	Raw pixels mouth	LSP of LPC	CNN/ Bi-GRU/FFNN	OuluVS2	PESQ	×	×	×	✓
	2018	Kumar et al. [39]	Raw pixels mouth	LSP of LPC	CNN/LSTM/FFNN	OuluVS2	PESQ	×	×	×	✓
	2019	Salik et al. [40]	Raw pixels mouth	LSP of LPC	STCNN/Bi-GRU	OuluVS2	PESQ	×	×	×	✓
	2019	Kumar et al. [41]	Raw pixels mouth	LSP of LPC	CNN/ Bi-GRU/FFNN	OuluVS2	PESQ	×	×	✓	✓
	2019	Uttam et al. [42]	Raw pixels mouth	AE features/ spectrogram	CNN/LSTM/FFNN/AE	OuluVS2	PESQ/Corr2D	×	×	✓	✓

2.2. Reconstructed Speech with Unconstrained Datasets

Recent research has aimed to improve reconstructed speech quality using datasets containing more speech from individual speakers without any constraints, such as Lip2Wav [29], AVSpeech [31], and UTKFace [31]. Prajwal et al. [16] introduced an adapted autoregressive sequence-to-sequence model, Lip2Wav, derived from Tacotron 2, to produce mel-spectrograms based on video frames. Their emphasis was on precise lip-to-speech mappings for individual speakers within the unconstrained Lip2Wav dataset, encompassing 120 h of talking face videos from five speakers. Through comparisons with prior research [21,23] in the same dataset, their approach resulted in a substantial fourfold increase in overall speech intelligibility.

Subsequent studies aimed to enhance the Lip2Wav model. He et al. [29] introduced a flow-based non-autoregressive lip-to-speech model named GlowLTS, designed to bypass autoregressive constraints and facilitate faster inference. Varshney et al. [30] used framed frame sequences as feature distributions using transformers within an autoencoder context. Their approach yielded marginal improvements, raising the Short-time Objective Intelligibility measure (STOI) measurements from 0.377 to 0.394 and 0.490, respectively, in the Lip2Wav dataset for Chemistry Lectures. In addition, Millerdurai et al. [31] presented the Lip2Speech method, with key design choices to achieve accurate lip-to-speech synthesis with the unconstrained datasets AVSpeech, LRW and UTKFace. They achieved STOI measurements of 1.38 with the LRW dataset and 3.55 MOS score compared to the groundtruth of 4.56 with the LRW dataset.

These strategies offer viable solutions for unconstrained lip-to-speech synthesis, aiming to enhance speech quality and intelligibility through large datasets. Nonetheless, their reliance on a two-stage pipeline involving the Griffin-Lim algorithm constrains audio quality and inference speed. Additionally, the speaker-specific training approach necessitates separate models for each speaker, entailing speaker dependence and restricting broader applicability.

2.3. Reconstructed Speech with Multiple Speakers

Regarding speaker-dependent models, Takashima et al. [26] introduced an exemplar-based strategy for multi-speaker reconstruction. They used a CNN to extract high-level acoustic features from visual frames, aiding the estimation of target spectrograms through an audio dictionary. Vougioukas et al. [23] proposed a GAN for directly estimating speech signals from mouth-focused video frames. While this enabled intelligible speech synthesis in a speaker-independent context, the PESQ score reached only 1.24 on unseen speakers with the GRID dataset. This is arguably due to the model's generation of raw waveforms, for which appropriate loss functions were challenging to ascertain.

Oneat et al. [18] introduced a video-to-speech architecture to enhance previous multi-speaker speech reconstruction methods. They integrated additional speaker-related input, including discrete identity or speaker embeddings, to separate linguistic content and speaker identity. Adversarial losses were introduced to segregate identity from video embeddings. Their model used both video and a specified identity as input, producing utterances in the chosen identity while preserving content intelligibility (WER 42.7%) and showcasing effective speaker control (EER 7.3%) across synthesised audio from videos of 14 diverse speakers. Similarly, Um et al. [19] proposed a multi-speaker face-to-speech waveform generation model, employing a GAN alongside linguistic and speaker characteristic features. Their approach enabled the direct conversion of face images into speech waveforms through end-to-end training. Linguistic features were extracted from lip movements via a lip-reading model, and speaker characteristics were predicted by face encoders through cross-modal learning. This model achieved MOS scores of 3.74 and 3.87 for seen and unseen data with four speakers from the GRID dataset. Wang et al. [20] proposed a multi-speaker VTS system based on cross-modal knowledge transfer from voice conversion (VC). Utilising vector quantization with contrastive predictive coding (VQCPC) for VC's content encoding, discrete phoneme-like acoustic units were derived. Their system

achieved PESQ scores of 1.816 and 1.417 on seen and unseen speaker tests with GRID speakers. While they achieved multi-speaker speech reconstruction to a certain extent, these studies all necessitate additional information such as speech or output text supplementary information to reconstruct speech from multiple speakers. In the absence of explicit speaker details, distinguishing and reconstructing speech from different speakers becomes challenging for these models. Moreover, their intricate structures hinder lightweight deployment and implementation on portable devices.

2.4. Multi-Task and Multi-View Speech Reconstruction

Qu et al. [34] introduced LipSound [34] and its successor, LipSound2 [35], which use visual information for speech synthesis and speech recognition for transcription. By capturing mouth movements, they reconstructed speech as mel-spectrograms, forming the basis for recognition. Speaker-dependent evaluations show that LipSound's mel-spectrograms exhibit a 0.843% character error rate and 2.525% word error rate with the GRID Dataset. LipSound2 achieves an average PESQ score of 1.72 with the GRID and TCD-TIMIT datasets.

Subsequent research [38–40,42] focused on multi-view speech reconstruction, aiming to enhance speech quality through the integration of visual inputs from multiple angles. Kumar et al. [39] employed multiple mouth angles to estimate the LSP representation of LPC coefficients for audio. Notably, they achieved an Average PESQ score of 2.5291 for Speaker 1 (Male) and 2.6255 for Speaker 10 (Female) from the OuluVS2 database. Uttam et al. [42] extended this to multi-view speaker-independent techniques by using multi-view speaker videos and lip pose angles to identify optimal views. A decision network selected the best view combination and model for speech signal reconstruction. Their results showed a PESQ score of 1.623 and a Corr2d score of 0.816 with the OuluVS2 database.

These studies aimed to enhance speech quality through text information output and multi-view utilisation. However, a key limitation is the lack of consideration for speaker independence. Additionally, their system structures are complex, demanding substantial data, resulting in high computational cost.

2.5. Review Conclusions

Although research has been conducted on speech reconstruction from silent videos, there are two crucial issues. Firstly, while ensuring the quality of speech prediction, the model's performance also needs to be considered. This includes factors such as the model being lightweight and robust, as these are essential for practical deployment. To the best knowledge of the authors, there is no relevant research addressing this problem in this domain, as the key focus of current research is on speech reconstruction with deep neural networks without considering lightweight solutions. Secondly, speaker-independent speech reconstruction results are currently unsatisfactory. Reconstructing speech from silent videos involving multiple speakers requires supplementary audio or textual inputs. There is a need for a direct speech reconstruction model reconstructing speech from multiple speakers without additional supplementary data.

3. Gabor-Based Visual Feature Extraction

Gabor filters are a widely used technique in image processing and computer vision, where an image is convolved with a bank of Gabor filters at different orientations and scales, resulting in a multi-channel representation of the image that captures both spatial and frequency information [43]. In the context of facial recognition, Dakin and Watt [44] used Gabor filters to identify human faces, and found that horizontally oriented features were the most informative, forming a distinctive "barcode" mapping of the face, with clear distinctions between features such as the lips, teeth, philtrum, and mentolabial sulcus.

This was extended in previous research by the authors [3] to capture more detailed mouth movement features with a novel Gabor feature extraction method for capturing 3D geometric lip features using Gabor-based image patches. This can effectively extract

three-dimensional features from speakers such as mouth opening area and depth, for a better understanding of mouth movement during speaking. Recent studies [6,11,45] have demonstrated the effectiveness of the Gabor feature extraction method in visual speech recognition tasks. The authors previously applied Gabor feature extraction for English and Mandarin Chinese speech recognition [11], which achieved comparable performance to deep learning CNN-based approaches while maintaining system simplicity and explainability. Compared to other visual extraction methods such as Active Appearance Models (AAM) [46], 2D-DCT [47], and CNNs [48], Gabor features strike a good balance between interpretability and accuracy, enabling parameters to be easily adapted and used for detailed analysis [3]. They also offer the advantage of being explainable and robust enough to recover from errors, making them suitable for real-world speech analysis applications such as speech recognition, synthesis, tracking, and linguistics. By first applying a Gabor transform and then extracting geometric features, information about the movement and shape of the lips during speech production can be visualised. Our Gabor feature extraction method was introduced in Li et al. [49], and is summarised below:

3.1. Proposed Feature Extraction Approach

Figure 1 illustrates the feature extraction flow chart, which involves tracking the ROI, filtering, and calculating Gabor lip region features, which will be explained in more detail below.

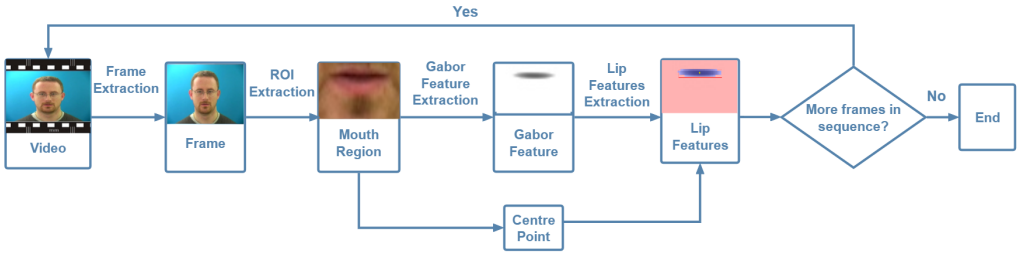


Figure 1. Flowchart of the lip feature extraction system showing frame extraction, ROI identification and extraction, Gabor transform, and lip feature extraction .

Gabor Lip Image Generation

Given a sequence of images I_n (where $n = 1 \ldots N$) extracted from a video file, the widely used Dlib toolkit [50], along with a trained model file, Shape-Predictor-68-Face-Landmarks, used in previous research by the authors [3,11], is used to identify 68 landmark face features. To detect the mouth region (ROI), four features are selected, represented by four x and y coordinate pairs $L_{nx}(1,2,3,4)$ and $L_{ny}(1,2,3,4)$, respectively, and $C_{nL}(X,Y)$, the centre point of the ROI (mouth region), is calculated as shown in Figure 2.

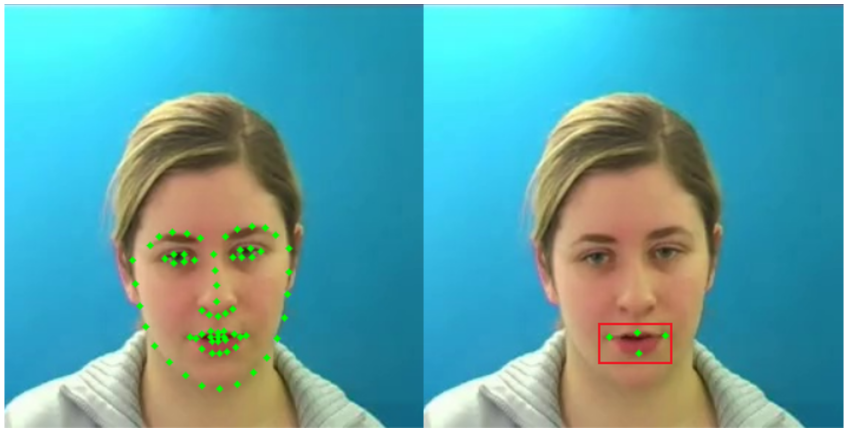


Figure 2. (**left**) Landmark features detected by the dlib toolkit, and (**right**) ROI extracted using the 4 landmark features.

Horizontal Gabor filters of greyscale images are then calculated, using the real component of a Fast Fourier Transform:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = exp(-\frac{x\prime^2 + \gamma^2 y\prime^2}{2\sigma^2}) \cos(2\pi \frac{x\prime}{\lambda} + \psi) \quad (1)$$

where:

$$x\prime = x \cos\theta + y \sin\theta$$

$$y\prime = -x \sin\theta + y \cos\theta$$

This is implemented using the Opencv Python function:

$$cv2.getGaborKernel((Ksize, Ksize), \sigma, \theta, \lambda, \gamma, \psi) \quad (2)$$

This has six parameters:
Ksize is the Gabor kernel size;
σ is the standard deviation of the Gaussian function used in the Gabor filter;
θ is the orientation of the normal to the parallel stripes of the Gabor function, which is 90 degrees;
λ is the sinusoidal factor wavelength in the equation;
γ is the spatial aspect ratio;
ψ is the phase offset, defined as 0 by default.

The above six parameters control the shape and size of the Gabor function. The role of each parameter is discussed in detail below. To illustrate the effects of parameters, the following values were chosen as a starting point: *ksize* = 50, θ = 90, λ = 10, σ = 5, γ = 0.5, ψ =0. Figure 3 is a GRID dataset image that is used here to demonstrate Gabor feature extraction in the following illustrations.

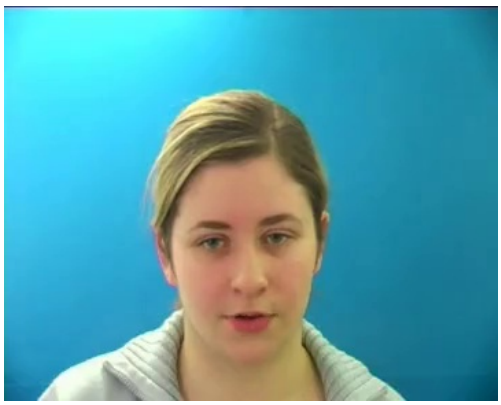
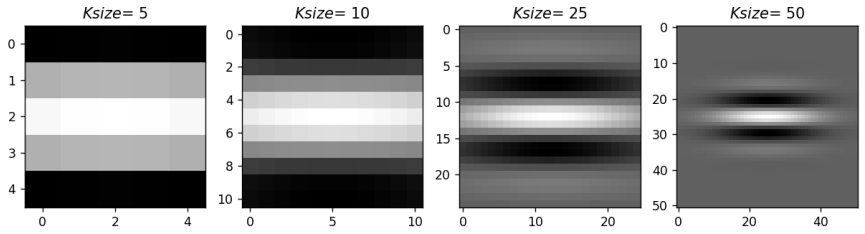


Figure 3. A sample image used to demonstrate Gabor feature extraction.

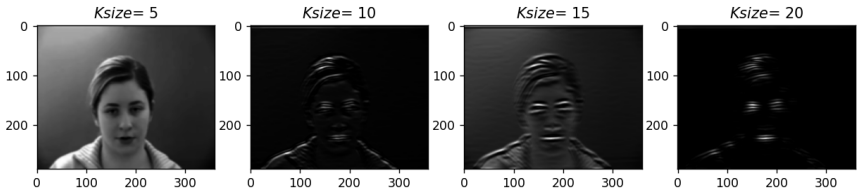
Ksize: When varying *ksize*, the size of the convolution kernel varies. Figure 4 shows example results when applying *ksize* with 5, 10, 25, 50 pixels. As shown in Figure 4b, when *Ksize* = 15, more efficient information about facial features can be extracted. That means a proper Gabor parameter is vital for feature extraction.

Wavelength (λ): The wavelength governs the width of the strips of the Gabor function. Increasing the wavelength produces thicker strips and decreasing the wavelength produces thinner strips. Keeping other parameters unchanged and changing the wavelength to 5, 10, 15, and 20, the stripes shown in Figure 5 get thicker.

Orientation (θ): Theta (θ) controls the orientation of the Gabor function. The zero-degree theta corresponds to the vertical position of the Gabor function. Figure 6 shows the different orientations of the Gabor kernel. Different orientations of the Gabor filter highlight different features of the face. In Figure 6b, the vertical filter (θ = 0) obtains more vertical information, such as ears and neck, whereas the Horizontal filter (θ = 90) receives information about the eyes and mouth. The 45 and 135-degree filters can highlight other information.

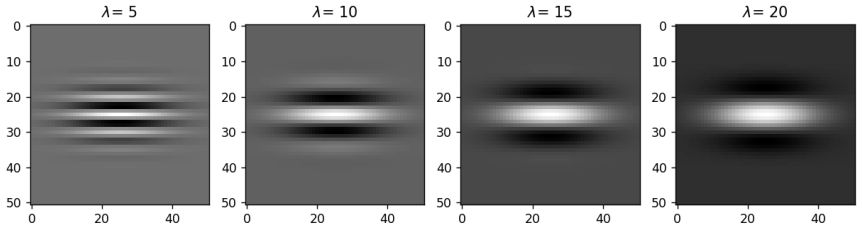


(**a**) The Gabor kernel with 5, 10, 25, 50 pixel kernel sizes.

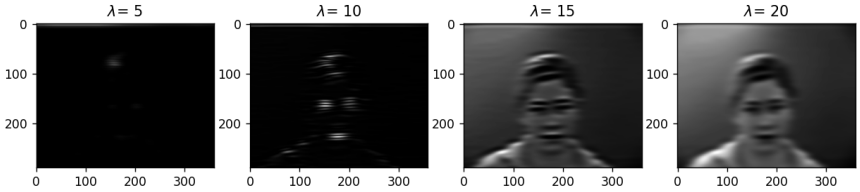


(**b**) Facial Gabor features with 5, 10, 25, 50 pixel kernel sizes.

Figure 4. Gabor kernels and facial Gabor features with different kernel sizes.

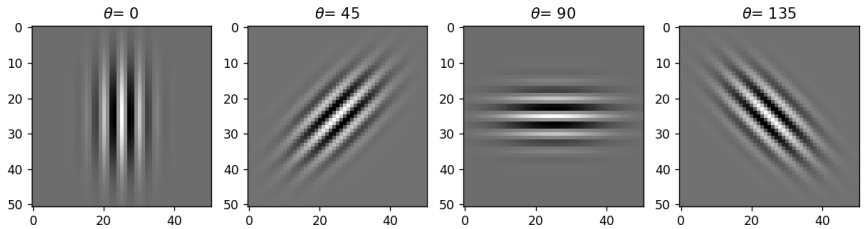


(**a**) The wavelength of the Gabor kernel with sizes of 5, 10, 25, 50.

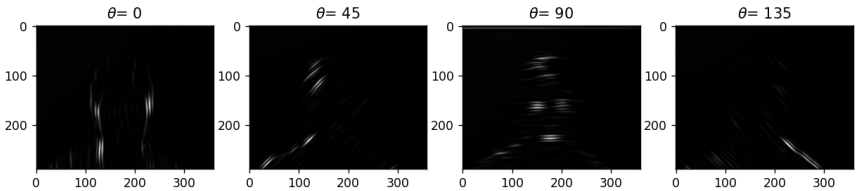


(**b**) The facial Gabor feature with 5, 10, 25, 50 kernel wavelengths.

Figure 5. The Gabor kernel and facial Gabor features with different wavelengths.



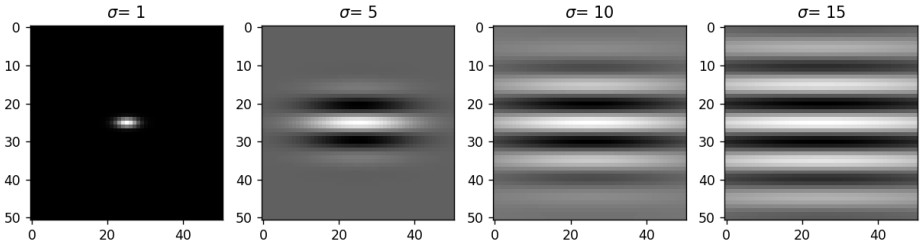
(**a**) The orientation of the Gabor kernel for 0, 45, 90, 135 degrees.



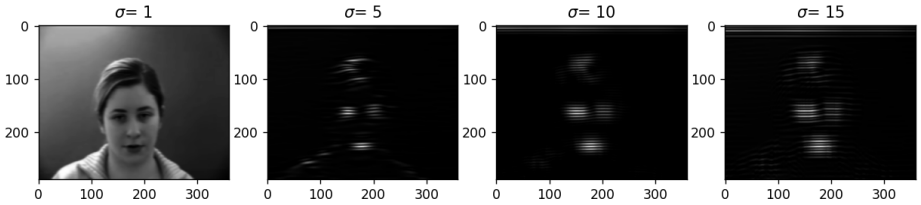
(**b**) The facial Gabor feature with 0, 45, 90, 135 degrees orientation.

Figure 6. The Gabor kernel and facial Gabor features with different orientations.

Standard deviation of the Gaussian envelope (σ): The sigma parameter (σ) controls the overall size of the Gabor envelope. For larger values of σ, the envelope increases, allowing more stripes, and with smaller sigma values, the envelope tightens. Figure 7 shows that with increasing sigma values from 1 to 15, the number of stripes in the Gabor function increases. In Figure 7b, it can be seen that when $\sigma = 15$, more stripes can be seen in the features of the eyes.



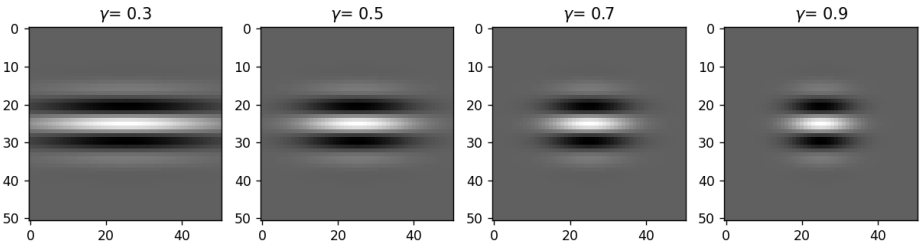
(**a**) The sigma of Gabor kernel with 1, 5, 10, 15.



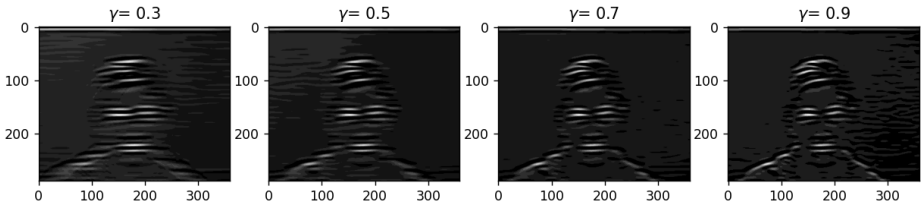
(**b**) The facial Gabor feature when sigma of 1, 5, 10, 15 is applied.

Figure 7. The Gabor kernel and facial Gabor features with different sigma values.

The spatial aspect ratio (γ): The gamma parameter (γ) controls the ellipticity of the Gaussian function. When $\gamma = 1$, the Gaussian envelope is circular. However, according to Jones and Palmer [51], it has been found to vary in a limited range of $0.23 < \gamma < 0.92$. As shown in Figure 8, when increasing the value of gamma from 0.3 to 0.9, keeping other parameters unchanged, the width of the Gabor function reduces. Figure 8b shows shorter eye and mouth features.



(**a**) The gamma of Gabor kernel with 0.3, 0.5, 0.7, 0.9.



(**b**) The facial Gabor features when applying sigma with 0.3, 0.5, 0.7, 0.9.

Figure 8. The Gabor kernel and facial Gabor features for different gamma values.

In this work, four parameters need to be adjusted: *Ksize*, σ, λ, and γ. Preliminary investigation identified the optimal parameters as: *Ksize* = 12, $\sigma = 6$, $\lambda = 15$, and $\gamma = 0.5$.

3.2. Visual Input I: Gabor Lip Image

To minimise noise, filtering is applied to the image with thresholding [52]. To determine the optimal segmentation threshold, the Python function filters.threshold_yen is used with the Yen algorithm [53]. This identifies each pixel in the transformed image as belonging to either the target or background region. The resulting image is used to create the resized Gabor Lip Image (GLI), a 28 × 28 2D image, as shown in Figure 9.

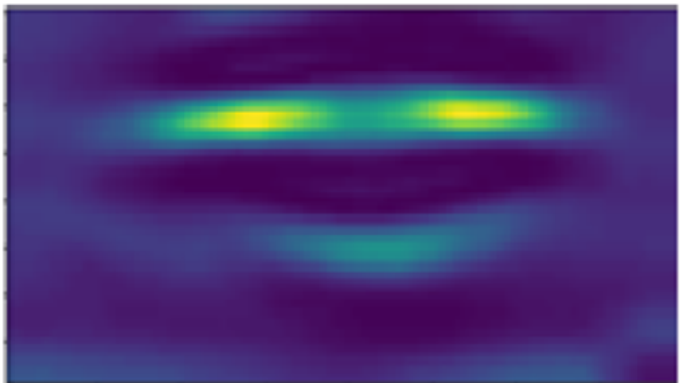


Figure 9. 2D Gabor Lip Image (2D-GIL).

Recent research [2,21,25,26,28] on speech reconstruction often uses raw pixel inputs of the entire face as the model input, but using the entire face results in a large amount of visual input data that can be very challenging for a model to learn. For example, in Akbari et al. [2], the input pixel size of each frame is (228, 228, 3), meaning that the number of pixel values that the model needs to learn in each frame is as high as 155,952. However, many of these values are not related to speech, which can affect the model's performance and robustness. By using the Gabor filtering method to remove arguably less relevant information and focus on the mouth region, the input size can be reduced from 155,952 to just 784 inputs. Furthermore, these filtered data values can more intuitively reflect the changes in lips when speaking.

3.3. Visual Input II: Gabor Lip Features

In addition to using Gabor Lip Images as network input, this paper also reports experiments with Gabor Lip Features. For each target lip image, the Gabor lip features are calculated. The Python *regionprops* function is used to identify the closest region *r* to the ROI centre point as the target lip region and calculate features in the form of a lip 'patch' covering the mouth area, as shown in Figure 10. The feature extraction approach has been fully described in previous research [49], and full details can be found there.

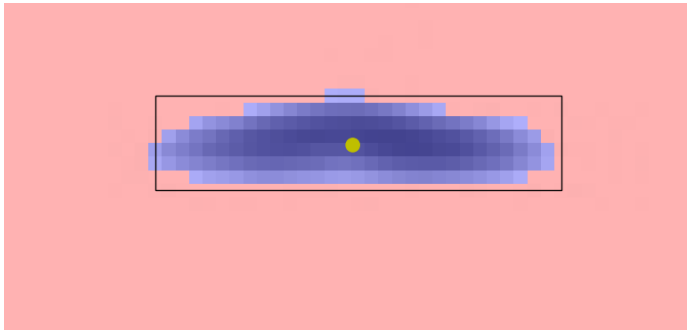


Figure 10. Gabor Transformed ROI.

For each lip patch *r*, seven feature values are generated: width, height, area, intensity, x and y values of the central point, and orientation. The box width is the inter-lip width. The height is the inter-lip height. The area is the number of pixels, The intensity is the sum

of each pixel density value (the darker the inter-lip area, the deeper the mouth opening and the larger the sound intensity), as shown in Figure 11.

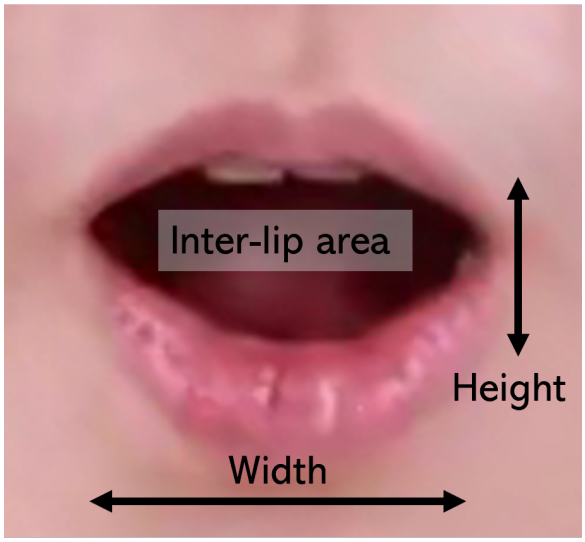


Figure 11. Example of mouth ROI, showing height, width, area, and intensity.

Compared to pixel-based inputs such as face pictures or Gabor Lip Images, Gabor lip features significantly reduce the quantity of visual input data required for speech reconstruction. The original pixel-based image input, which required 155,952 inputs (one for each pixel), can be reduced to just seven Gabor lip eigenvalues, resulting in a significantly smaller network input. By utilising Gabor lip eigenvalues, which directly reflect lip changes, the number of layers in the model can be reduced by removing all CNN layers entirely, resulting in faster speech reconstruction and more lightweight models.

4. Gabor-Based Speech Reconstruction Network

Gabor lip features offer valuable insights into lip shape and motion during speech production, suggesting the theoretical feasibility of generating speech signals from lip movements. In this paper, two novel speech reconstruction systems are introduced, based on distinct visual features. The first model, GaborCNN2Speech, enhances the baseline LipAudSpec model. It replaces mouth image inputs with Gabor lip images to simplify the input, and reduces the number of CNN layers from 7 to 1. The CNN output is then fed into an LSTM network to capture temporal sequence information, generating an autoencoder spectrogram, which is subsequently converted into an audio signal. The second model, GaborFea2Speech, uses Gabor features as inputs, further reducing input complexity and eliminating all CNN layers. The features are directly input into an LSTM network to produce the autoencoder spectrogram. Figure 12 shows the overall model architecture, where (i) presents the Lip2AudSpec baseline model, which is the CNN-based model used by Akbari et al. [2], and is used here for comparison; (ii) displays our proposed GaborCNN2Speech model, which is fully introduced in Section 4.1; and (iii) illustrates our proposed novel GaborFea2Speech which uses the Gabor feature extraction presented in Section 3.3 as the input to the simplified model, which will be introduced in Section 4.2. In this work, an LSTM-based model has been chosen for several reasons: Firstly, it mirrors the approach of Akbari et al. [2], who use an LSTM for their baseline model, and allows for a consistent comparison. In addition, previous research by the authors has identified that LSTM-based approaches can achieve good results when using Gabor features for speech recognition [6,11], and so this approach is used in this paper.

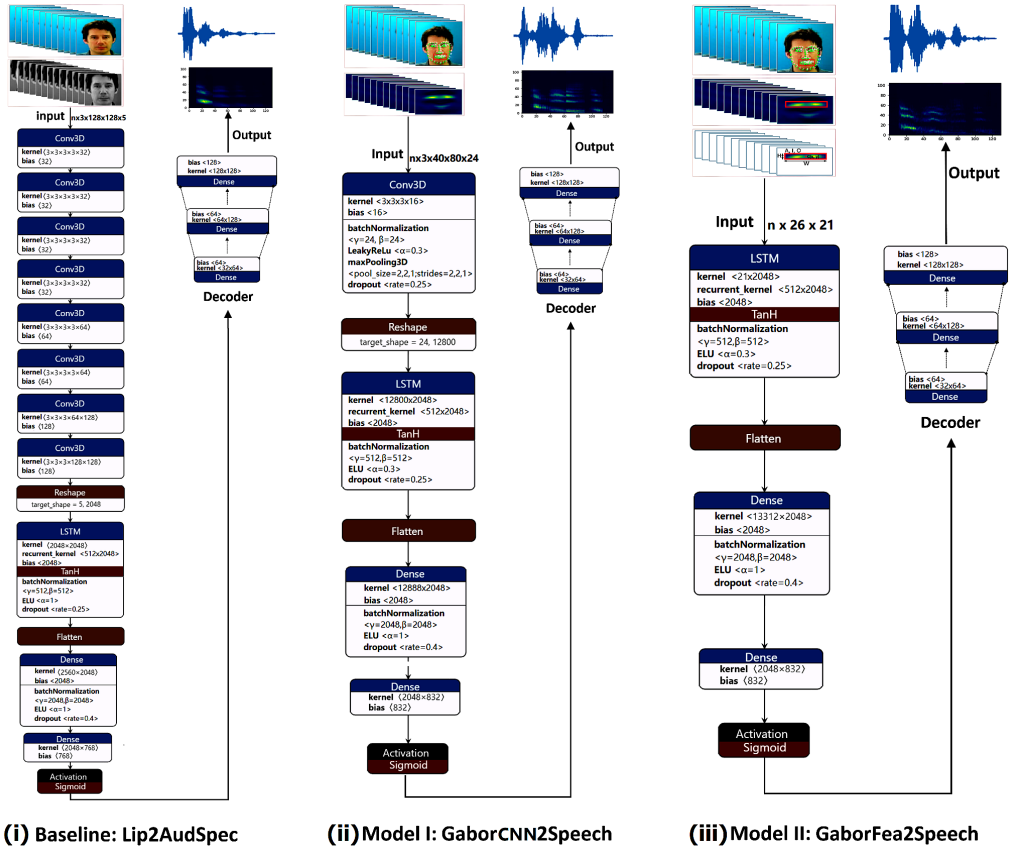


Figure 12. Architecture of Gabor-based speech reconstruction networks. (**i**) presents the Lip2AudSpec baseline model, (**ii**) displays our proposed GaborCNN2Speech model, and (**iii**) illustrates our new GaborFea2Speech model.

4.1. Model I: GaborCNN2Speech

Our GaborCNN2Speech model represents a novel approach to speech reconstruction, employing a Gabor CNN-based architecture that harnesses the texture and edge information extraction capabilities of Gabor filtering and the robust image recognition abilities of CNNs. This integration of techniques seeks to enhance visual feature extraction performance while simplifying model complexity. Gabor filtering is known for its proficiency in extracting robust and discriminative image features, which are subsequently effectively processed by CNNs to recognize intricate patterns and structures. This combined approach has demonstrated high accuracy in various domains, including visual recognition [54], speech recognition [55], and emotion recognition [56].

In preliminary experiments (not reported here), the architecture design of the GaborCNN2Speech model was explored by reducing the number of CNN layers and parameters (thus reducing the complexity). We then compared the training time, the inference time, and the accuracy of the spectrogram reconstruction. Therefore, gradually reducing the number of CNN layers and model parameters from seven layers to one layer while simultaneously decreasing its parameters was trialed, thus reducing the training time and consumption of memory and energy. It was found that good results could be delivered with just a single CNN layer, and this structure is used for the results presented in this paper.

GaborCNN2Speech Model Structure

Our proposed GaborCNN2Speech model is shown in Table 2. The input is a sequence of Gabor mouth pictures extracted from each frame and resized to have dimensions $W \times H$. It is then divided into K non-overlapping slices each of length L_v. First, and second-order temporal derivatives at each frame were calculated to form a 4D tensor of shape (3, H, W, L_v), where 3 is the number of time-derivative channels (0th, 1st and 2nd order). The target

bottleneck feature vector was also divided into K slices with length L_v and no overlap. A 1 layer convolutional network followed by a 3D max pooling layer is used to extract spatial and temporal features of the video sequence. The CNN maintains its spatial dimension and order so that an LSTM can model time dependencies, so its output is reshaped to a tensor of shape (L_v, N_f), in which N_f represents the spatial features extracted by the convolutional network. This reshaped tensor is fed into a single-layer LSTM network with 512 units to capture the temporal pattern. The output of this layer is further flattened and fed into a single-layer fully connected network and then finally to the output layer. The output layer has 32*L_a units to give 32-bin * L_a-length bottleneck features which are then connected to the decoder part of the pre-trained autoencoder to reconstruct the auditory spectrogram. The audio waveform is then reconstructed using NSRTools, the auditory spectrogram toolbox developed by Chi et al. [57].

Table 2. Structure of GaborCNN2Speech Network.

Layers	Size
Input Layer	(None, 3, 28, 28, 5)
Conv3D (32) BatchNormalization () LeakyReLU MaxPooling3D (2, 2, 1) Dropout (0.25)	(None, 16, 40, 80, 5)
Reshape	(None, 5, 12,800)
LSTM (512) BatchNormalization () ELU (alpha = 1.0) Dropout (0.25)	(None, 5, 512)
Flatten	(None, 2560)
Dense (2048) BatchNormalization () ELU (alpha = 1.0) Dropout (0.4)	(None, 2048)
Output Dense (832) Activation (sigmoid = 0.05)	(None, 832)

4.2. Model II: GaborFea2Speech

Our proposed GaborFea2Speech model uses Gabor Lip Features (as detailed in Section 3.3) as inputs. Instead of the original 28 × 28 pixel Gabor lip images, seven key visual features (width, height, mass, area, Xpos, Ypos, and orientation) are used to represent lip dynamics. This reduction in input data dimensionality, from 782 to 7, is accompanied by an enhancement of the feature set through the incorporation of three-time derivatives, effectively extending the Gabor features. This extended Gabor feature set serves as a direct visual representation of lip changes during speech, eliminating the need for complex CNN layers. The preprocessed Gabor features are directly fed into a 1-layer LSTM sequence network to generate the corresponding audio spectrogram. To the best knowledge of the authors, this is the first attempt to use Gabor Features for speech reconstruction, and this approach offers advantages in terms of lightweight structure and speed of training. Moreover, the objective is to harness the distinctness of Gabor feature visual attributes across different speakers by developing a multi-speaker speech reconstruction model based on Gabor Features, thereby achieving greater speaker independence.

Pre-processing is an essential step for GaborFea2Speech. In each 3-second Grid video, the Gabor features input matrix has dimensions $F_n * L_f$ (7 × 75), where F_n represents the feature dimensions and L_f is the frame length. The output is a spectrogram with

dimensions $S_n * L_s$ (32 × 390), where S_n represents the encoded spectrogram components and L_s denotes the total spectrogram time frequency. When reconstructing speech using Gabor features, the main challenge is the mismatch between the shorter input length of 75 and the longer output length of 390, which can result in inaccurate predictions and information loss. Moreover, the lower input feature dimension of 7 as compared to the desired output dimension of 32 may result in reduced accuracy and difficulties in capturing complex relationships in the data. We experimented with pre-processing to investigate different input sequence lengths and feature dimensionalities.

4.2.1. GaborFea2Speech-Input Sequence Dimension Testing

The effect of input sequence length on speech reconstruction can be evaluated by extracting input sequences with different numbers of frames from the original videos. The *write_video* function from the *movie.py* package was used to extend the input sequences during the pre-processing stage of the GaborFea2Speech model. By adjusting the frame extraction rate, the original rate of 25 fps was transformed to 85 and 130 fps, which increased the frame count of each video from 75 to 255 and 390 frames, respectively, where 390 matches the desired output sequence length. 100 videos from Speaker 1 were selected, the accuracy and PESQ scores of the predicted spectrograms was evaluated (more detail will be provided in the main experiments section), and a sample was randomly selected and its spectrogram and speech waveform plotted to provide a visual assessment of different input sequence lengths. These preliminary results indicated that the matching of input and output sequences is a critical factor in the success of our GaborFea2Speech model. If the input sequence is only 75 frames, significantly shorter than the output sequence, the model cannot accurately predict the desired output, which results in unintelligible speech. With the increase in input sequence length to 255 frames, which is still short of the output sequence, the model captures a significant portion of the input information, resulting in higher prediction accuracy and quality than with 75 frames. Upon extending the input sequence to 390 frames, which matches the length of the output sequence of the model, the model successfully predicts all essential speech information in the preliminary trials, resulting in improved prediction accuracy and PESQ scores. Therefore, this configuration is used in the experiments reported in the following sections.

We also assessed whether input dimensionality affected the quality of reconstructed spectrograms and speech by varying input dimensionality while keeping the input sequence length constant at 390. By applying the time derivative method to input features, it was possible to increase the dimensionality of the input features in order to ensure that they accurately represent the data and capture the necessary information. With the addition of 1D, 2D, 3D, and 4D time derivative channels, the 7 Gabor features were expanded to 14, 21, 28, and 35. It was found that the accuracy of reconstructed spectrograms improves with increasing the number of input features. The highest accuracy was achieved by using 21 features. As a result of these preliminary trials, the GaborFea2Speech model is configured to use 2-D time derivatives and is also interpolated to have an input sequence of 390 frames.

4.2.2. GaborFea2Speech Model Structure

The final structure of our proposed GaborFea2Speech model is shown in Table 3. Input to the network is a sequence of Gabor features, denoted as F_n, extracted from silent video frames. To increase the length of the input sequence, video frames are upsampled to 130 fps, resulting in a 3-s sample video of 390 frames. This length is then divided into K non-overlapping slices L_v. Gabor features are augmented with first and second-order temporal derivatives to enrich the input representation, tripling the dimension of Gabor features. This fusion results in a 2D tensor with dimensions (L_v, 3*F_n). The target bottleneck spectrogram feature vector, which acts as the desired output, is also partitioned into K slices of length L_a without overlap. By reserving the spatial aspect of the preprocessed input, an LSTM can be directly used to model time dependencies. A single-layer LSTM network

with 512 units is utilised to capture temporal patterns. LSTM outputs are then flattened and passed through a single-layer fully connected network before reaching the final output layer. A similar procedure is followed in the output process as with the GaborCNN2Speech model. A sigmoid nonlinearity is used at the output layer, while the LSTM and fully connected layers use ELU activation functions. The audio waveform is then reconstructed using the auditory spectrogram toolbox [57], which estimates the phase that corresponds to the magnitude of the auditory spectrogram and performs an inverse transformation.

Table 3. Structure of GaborFea2Speech Network.

Layers	Size
Input Layer	(None, 26, 21)
LSTM (512) BatchNormalization () ELU (alpha = 1.0) Dropout (0.25)	(None, 5, 512)
Flatten	(None, 2560)
Dense (2048) BatchNormalization () ELU (alpha = 1.0) Dropout (0.4)	(None, 2048)
Output Dense (832) Activation (sigmoid = 0.05)	(None, 832)

5. Experiment Design

5.1. Dataset

The widely used [2,3,5,58,59] GRID audio-visual corpus [60] is utilised for training speech reconstruction models, allowing comparisons with other related research. The GRID corpus includes audio and video recordings of 34 speakers, each with 1000 utterances, with each utterance consisting of a combination of six words categorised into six categories from a 51-word vocabulary.

Three tests were conducted: individual speakers, multiple speakers, and vocabulary tests. To assess the generated speech on each individual speaker model, two male speakers (S1, S27) and two female speakers (S16, S25) were randomly selected. For the multiple speaker test, data from multiple speakers was combined, using four males (S1, S10, S17, S27) and four females (S11, S15, S16, S25). These resulted in larger datasets, using the combined 1000 sentences from each speaker. The dataset configurations are shown in Table 4. For the vocabulary test, 51 words were extracted from sentence samples (S1 and 100 samples) with the time-aligned file in order to train our proposed system on each word.

Table 4. Multi-Speaker train and test configurations.

Model	Speakers on Training Set (80%)	Speakers on Test Set (20%)
1-S	S1 (male) S25 (female) S16 (female) S27 (male)	S1 S25 S16 S27
2-S	S1, S27 (male ×2) S16, S25 (female ×2)	S1, S27 S16, S25
4-S	S1, S25, S16, S27 (male ×2, female ×2)	S1, S25, S16, S27
6-S	S1, S17, S27, S16, S15, S25 (male ×3, female ×3)	S1, S17, S27, S16, S15, S25
8-S	S1, S10, S17, S27, S16, S11, S15, S25 (male ×4, female ×4)	S1, S10, S17, S27, S16, S11, S15, S25

5.2. Implementation

Our system used a GeForce GTX 1050ti GPU (16GB of memory). Keras [61] (with TensorFlow backend) was used in Python 3. NSRtools [57] in MATLAB was used to transform auditory spectrograms and waveforms.

For GaborCNN2Speech, the length of each video slice L_v was set to 5, and the length of each audio slice L_a to 26, to ensure that there are an equal number of audio and video slices K. With Gabor lip pictures cropped with a width W and height H of 28, enough lip features could be extracted. For GaborFea2Speech, seven Gabor features (F_n) were extracted from each video frame and 2-time derivative channels were added to extend the feature number to 21. The length of each video slice L_v is 26, which equals the length of each audio slice L_a with the same number of audio and video slices K. Both the visual and audio data were divided into training, validation, and test sets, with ratios of 80%, 10%, and 10%, respectively.

For model training, weight initialization was performed using the method by proposed by He et al. [62]. Batch normalization [63] was used for all layers, dropout [64] of $p = 0.25$ in the convolutional block, and the L2 penalty multiplier was set to 0.0005 for the convolutional layer. For LSTM and MLPs after the convolutional blocks, a dropout of $p = 0.3$ was used, and regularization was not used. The model was trained using a batch size of 15, and the parameter α for ELU non-linearity [65] was set to 1. To improve the robustness of the network, data augmentation was performed in each epoch by randomly selecting videos and either flipping them horizontally or adding a small level of Gaussian noise. This approach mirrors the methodology of Akbari et al. [2], who identified that, although the reconstruction performance dropped slightly, the overall lip performance structure improved significantly. This paper, therefore, used the same methodology. Optimization was performed using Adam [66] with an initial learning rate of 0.0001, which was reduced by a factor of 5 if the validation loss was not improving for 4 consecutive epochs. The loss function for all networks was CorrMSE [2]. For auditory spectrogram generation, NSRtools was used, a Matlab toolbox by Chi et al. [57]. Parameters for all auditory spectrogram generation and audio waveform reconstruction from the spectrogram were $frm\_len = 10$, $tc = 10$, $fac = -2$, and $shft = -1$.

5.3. Metrics

Several well-established metrics were used to evaluate the accuracy, intelligibility, and quality of the reconstructed speech. 2D Pearson's correlation coefficient (Corr2D) [67] was used to measure the correlation between the reconstructed and main spectrograms (ranging from -1 to +1). To assess the intelligibility, the Short-time Objective Intelligibility measure (STOI) was utilised, which has a range between 0 and 1. For evaluating the overall quality, Perceptual Evaluation of Speech Quality (PESQ) [68] was employed. PESQ returns a score between -0.5 and +4.5. Additionally, the Objective Overall Quality (OOQ) metric was utilised, introduced by Hu and Loizou [69], which combines individual objective measures such as PESQ, Log-Likelihood Ratio (LLR) [70], and Weighted-Slope Spectral Distance (WSS) [69], as defined in Equation (3):

$$C_Q = 1.594 + 0.80PESQ - 0.512LLR - 0.07WSS \quad (3)$$

6. Results

Here, results are reported using both single and multi-speaker models, as well as individual vocabulary tests, using the datasets introduced in Section 5.1. We compare our proposed GaborCNN2Speech and GaborFea2Speech models with a leading CNN-based approach, Lip2AudSpec.

6.1. Single Speaker Model Results

Four speakers were selected (see 1-S in Table 4), using 1000 sentences for each speaker. Our proposed networks were trained with an 80%, 10%, and 10% split for training, val-

idation, and testing, respectively. Corr2D, PESQ, OOSQ, and STOI were applied for evaluation, as discussed in Section 5.3. Additionally, to provide a visual comparison of the reconstructed results, one sample was selected from each speaker and the reconstructed spectrogram compared with the original. The results are shown in Table 5.

Table 5. Comparative evaluation of GaborFea2Speech and GaborCNN2Speech Networks against Lip2AudSpec for individual speaker speech reconstruction.

Measure	Model	S1	S16	S25	S27	AVG	IQR
Corr2D	GaborFea2Speech	60.90%	70.10%	65.13%	60.54%	64.17%	24.81%
	GaborCNN2Speech	63.60%	**70.80%**	**66.03%**	**64.70%**	**66.28%**	**14.18%**
	Lip2AudSpec	**64.40%**	69.10%	56.33%	54.90%	61.18%	17.25%
PESQ	GaborFea2Speech	1.488	**1.882**	1.617	**1.889**	**1.719**	0.624
	GaborCNN2Speech	1.399	1.816	**1.735**	1.805	1.689	0.550
	Lip2AudSpec	**1.646**	1.735	1.378	1.746	1.626	**0.536**
OOSQ	GaborFea2Speech	−0.334	**0.667**	0.402	**0.391**	0.282	0.848
	GaborCNN2Speech	−0.366	0.659	**0.566**	0.338	**0.299**	0.836
	Lip2AudSpec	−**0.145**	0.430	−0.039	0.037	0.071	**0.522**
STOI	GaborFea2Speech	**0.149**	**0.332**	**0.335**	0.226	**0.261**	0.208
	GaborCNN2Speech	0.147	0.313	0.315	**0.249**	0.256	0.170
	Lip2AudSpec	0.118	0.327	0.295	0.204	0.236	**0.159**

Table 5 compares GaborFea2Speech, GaborCNN2Speech, and Lip2AudSpec. While there are fluctuations for individual speakers, the average shows that our proposed GaborFea2Speech and GaborCNN2Speech models demonstrate slightly improved performance compared to Lip2AudSpec across all measures. GaborCNN2Speech achieves the highest scores in Corr2D (66.28) PESQ (1.689), and OOSQ (0.299), while GaborFea2Speech outperforms GaborCNN2Speech in the STOI measure (0.261 vs. 0.256) and Lip2AudSpec only achieves 0.236. This demonstrates that the proposed models better preserve speech intelligibility and similarity to the original speech. Regarding OOSQ, GaborCNN2Speech achieves the highest average score of 0.299, followed by GaborFea2Speech with 0.282. Lip2AudSpec obtains the lowest average score of 0.071. This is the largest difference between models and indicates that the proposed models effectively reduce speech distortion and enhance objective speech quality. The IQR values highlight the variation in model performance across different measures. GaborCNN2Speech generally demonstrates a more consistent performance based on lower IQR values, suggesting its stability across the evaluated measures. Conversely, GaborFea2Speech and Lip2AudSpec exhibit larger IQR values, indicating more dispersion in their results.

Figure 13 shows examples of reconstructed spectrograms. Figure 13a(1)–a(3) shows the original spectrograms of four selected speakers (S1, S25, S27) along with their corresponding sample sentences: “bin blue by f 8 please”; “bin green at l q now”; “bin blue with x 8 now”; “bin blue by p 8 now”. Figure 13b(1)–b(3) shows reconstructed GaborFea2Speech spectrograms, Figure 13c(1)–c(3) shows the spectrograms reconstructed with GaborCNN2Speech. For comparison, Figure 13d(1)–d(3) shows equivalent Lip2AudSpec spectrograms. The comparison clearly shows that our proposed models outperform the baseline model in terms of spectrogram reconstruction. Our spectrograms exhibit enhanced clarity and sharper spectral peaks, indicating a higher fidelity in reproducing the original speech characteristics. Notably, our GaborFea2Speech method demonstrates superior accuracy in reconstructing the frequency space and maintaining the temporal continuity of phonemes across window samples. It also demonstrates superior accuracy in reproducing the original speech characteristics, especially for capturing high-frequency information. In contrast, the baseline model struggles to retrieve high-frequency information.

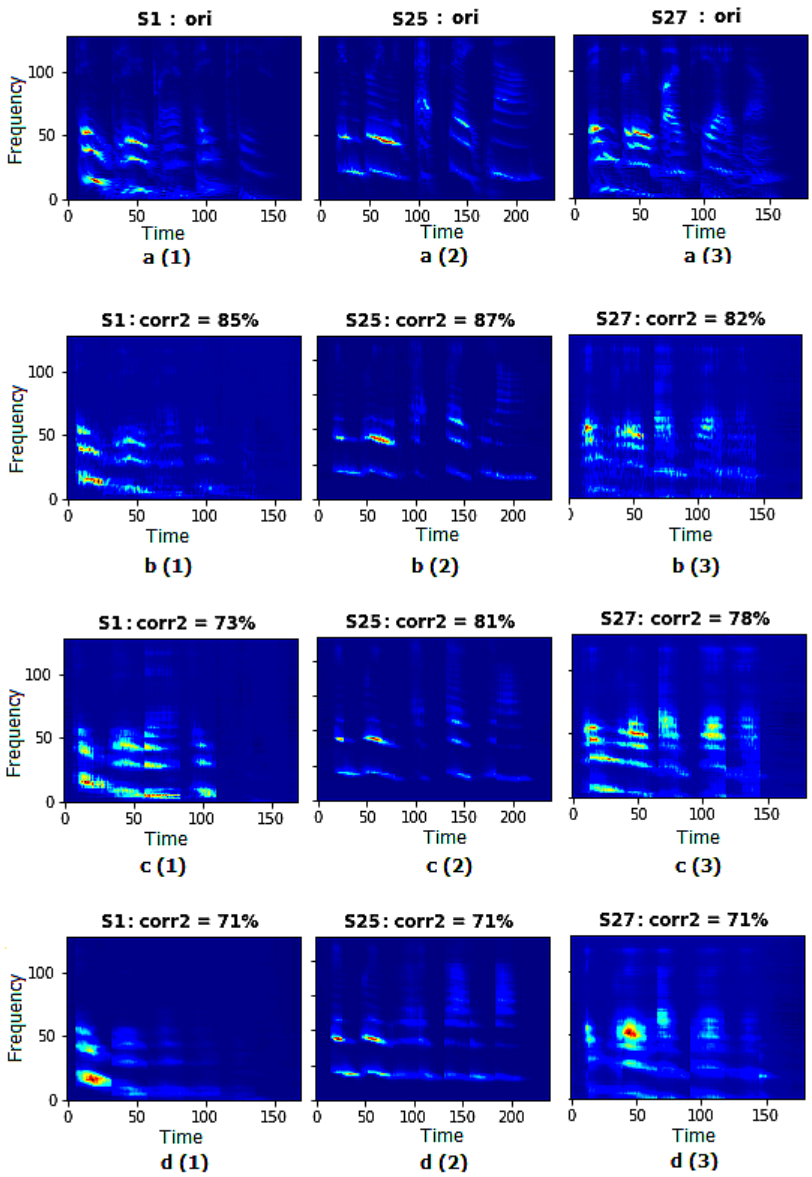


Figure 13. Comparative analysis of spectrogram reconstruction using proposed and baseline models. (**a(1)**–**a(3)**) show the original spectrograms. Reconstructed spectrograms from GaborFea2Speech and GaborCNN2Speech are displayed in (**b(1)**–**b(3)**) and (**c(1)**–**c(3)**), respectively. (**d(1)**–**d(3)**) show the Lip2AudSpec spectrograms.

6.2. Multi-Speaker Model Results

Multi-speaker speech reconstruction is important for evaluating model independence. Our proposed models are compared with the baseline, using the datasets described in Section 5.1, labelled as 1-S, 2-S, 4-S, 6-S, and 8-S. To validate performance, 100 samples from speaker S27 were randomly selected and the reconstructed speech from this speaker evaluated in various multi-speaker scenarios. The results are shown in Table 6 and Figure 14.

Table 6. Speech reconstruction results from multiple speaker models.

Measure	Lip2AudSpec					GaborCNN2Speech					GaborFea2Speech				
	1-S	**2-S**	**4-S**	**6-S**	**8-S**	**1-S**	**2-S**	**4-S**	**6-S**	**8-S**	**1-S**	**2-S**	**4-S**	**6-S**	**8-S**
Corr2D	60.70%	39.22%	39.28%	33.05%	29.55%	**66.18%**	**64.67%**	51.47%	49.91%	46.18%	64.77%	61.26%	**63.00%**	**58.17%**	**54.52%**
PESQ	1.627	1.414	1.458	1.476	1.431	1.696	1.697	1.591	1.703	**1.699**	**1.739**	**1.714**	**1.809**	**1.778**	1.684
OOSQ	0.071	−0.155	−0.304	−0.170	−0.384	0.300	0.243	−0.049	0.156	−0.042	**0.302**	**0.252**	**0.221**	**0.198**	**−0.031**
STOI	0.242	0.213	0.244	0.260	0.251	0.269	0.283	0.285	0.339	0.315	**0.278**	**0.287**	**0.330**	**0.355**	**0.320**

Table 6 shows that GaborFea2Speech consistently outperforms Lip2AudSpec and GaborCNN2Speech in all speaker scenarios. Under the Corr2D measure, GaborCNN2Speech achieves the highest percentage (66.18%) in the 1-S scenario, followed by GaborFea2Speech (64.77%). However, in the 2-S scenario, GaborFea2Speech outperforms both Lip2AudSpec and GaborCNN2Speech with a percentage of 61.26%. In terms of PESQ, GaborFea2Speech achieves the highest overall score (1.739), closely followed by GaborCNN2Speech (1.696). GaborFea2Speech consistently outperforms the other methods in terms of OOSQ, exhibiting the highest values across all scenarios. Similarly, under the STOI measure, GaborFea2Speech demonstrates the highest scores, indicating superior speech intelligibility compared to Lip2AudSpec and GaborCNN2Speech.

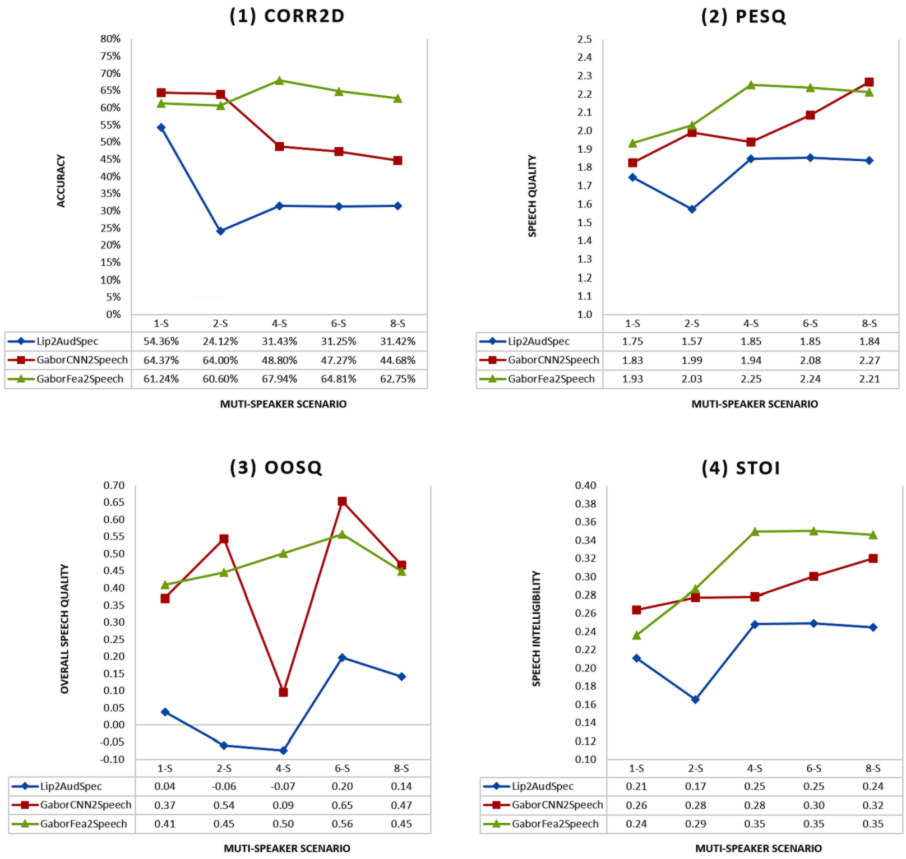


Figure 14. Comparative analysis of speech reconstruction for Speaker S27: GaborFea2Speech and GaborCNN2Speech models versus Lip2AudSpec with different datasets(1-S, 2-S, 4-S, 6-S, 8-S). Subplots (1)–(4) show the average accuracy (Corr2D), PESQ score, objective overall speech quality (OOSQ), and STOI measurement of the reconstructed speech

Figure 14(1–4) shows the average accuracy (Corr2D), PESQ scores, objective overall speech quality (OOSQ), and STOI measurement of the reconstructed speech for speaker S27 using models trained with different numbers of speakers. The results show the superior performance of the proposed GaborFea2Speech and GaborCNN2Speech models. Notably, the GaborFea2Speech model consistently outshines the other models across all metrics and scenarios. The results show that with all datasets and metrics, our GaborFea2Speech model delivers not only the best performance with multiple speakers but is also notably more stable with different datasets. There is a consistent pattern of both our proposed models outperforming the CNN model, even when trained with large datasets.

To further compare performance, a sample sentence was randomly selected ("bin blue by f 8 now") from speaker S27 and the spectrogram and waveform analysed to visualise the change across various multi-speaker scenarios. Figure 15i(0) shows the original spectrogram and waveform. Figure 15a(1)–a(5) presents the reconstructed spectrograms

and waveforms using our proposed GaborFea2Speech model in 1-S, 2-S, 4-S, 6-S, and 8-S scenarios, Figure 15b(1)–b(5) shows the equivalent for our proposed GaborCNN2Speech model, and Figure 15c(1)–c(5) shows Lip2AudSpec visualisations.

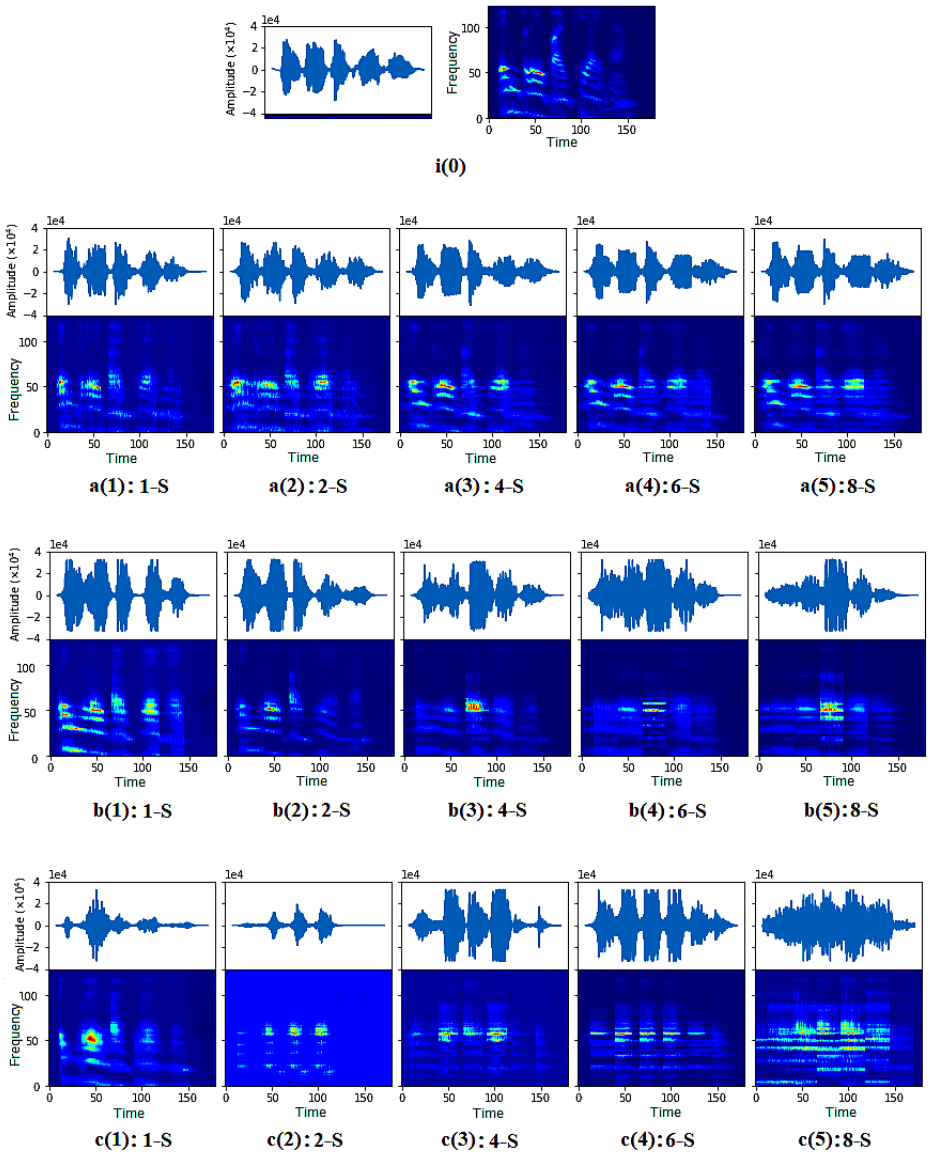


Figure 15. Reconstructed spectrogram and waveform comparison for Speaker S27. **i(0)**: original spectrogram and waveform of "bin blue by f 8 now". (**a(1)–a(5)**): reconstructed spectrograms and waveforms for GaborFea2Speech with 1-S, 2-S, 4-S, 6-S, 8-S datasets. (**b(1)–b(5)**): equivalent GaborCNN2Speech results. (**c(1)–c(5)**) equivalent Lip2AudSpec results.

Figure 15 demonstrates that with our multi-speaker dataset, the proposed GaborFea2Speech model exhibits minimal sensitivity to the number of speakers when reconstructing spectrograms and waveforms, meaning that it is more suitable for multi-speaker speech reconstruction. GaborCNN2Speech is moderately affected by the number of speakers, but the baseline Lip2AudSpec model is heavily influenced by the number of speakers and is unable to reconstruct multi-speaker speech.

Figure 15a(1)–a(5) shows that GaborFea2Speech consistently and accurately reconstructs a significant portion of frequency content with excellent temporal continuity and phonetic accuracy as the number of speakers increases. Moreover, the reconstructed waveforms from GaborFea2Speech showcase high amplitude precision and minimal sensitivity to the number of speakers. GaborCNN2Speech exhibits partial reconstruction of frequency

content in spectrograms and amplitude information in waveforms when the number of speakers is low (Figure 15b(1)–b(3)). However, as the number of speakers increases, its reconstruction performance gradually declines. At six speakers, GaborCNN2Speech only captures the temporal continuity of phonetic samples, losing a significant amount of frequency content and amplitude information (Figure 15b(4)). The accuracy further decreases with eight speakers (Figure 15b(5)). In contrast, the Lip2AudSpec results in Figure 15c(1)–c(5) exhibit the poorest performance in multi-speaker speech reconstruction. It should be noted that with eight speakers, the model loses almost all frequency content and amplitude information, resulting in effectively random reconstruction of noise-related information.

6.3. Individual Word Results

In addition to complete sentences, GaborFea2Speech and GaborCNN2Speech performance with individual words was compared. The vocabulary from 1000 sentences of Grid Speaker 1 was extracted, which encompassed 6 categories and 51 distinct words, representative of the full set of vocabulary found in the Grid Corpus. The words were trained using GaborFea2Speech and GaborCNN2Speech models. Table 7 provides a comparison of GaborFea2Speech and GaborCNN2Speech across different word categories. The results indicate that GaborFea2Speech consistently outperforms GaborCNN2Speech in terms of accuracy, speech quality, overall quality, and intelligibility. Notably, adverbs were found to have the highest accuracy and speech quality, while letters had the best speech intelligibility based on GaborFea2Speech. GaborCNN2Speech demonstrates competitive performance in the command category for accuracy and the colour category for perceptual quality and intelligibility. These results highlight GaborFea2Speech's superior performance in various aspects of word reconstruction. This indicates its effectiveness in capturing and reconstructing speech from different word categories. This is also shown in the radar graph in Figure 16.

Table 7. Speech reconstruction from vocabulary: evaluating GaborFea2Speech and GaborCNN2Speech Networks on word category.

Measure	Model	Command	Colour	Preposition	Letter	Digit	Adverb	Average
Corr2D	GaborFea2Speech	79.15%	79.15%	76.15%	77.25%	80.50%	**80.56%**	**78.79%**
	GaborCNN2Speech	**73.32%**	72.85%	67.03%	69.52%	71.0%	73.01%	71.08%
PESQ	GaborFea2Speech	**1.32**	1.27	1.27	0.86	1.23	1.28	**1.21**
	GaborCNN2Speech	0.80	**0.82**	0.75	0.75	0.78	0.81	0.79
OOSQ	GaborFea2Speech	0.29	0.31	0.26	0.26	0.34	**0.45**	**0.32**
	GaborCNN2Speech	0.27	0.27	0.26	0.27	0.31	**0.3**	0.29
STOI	GaborFea2Speech	0.35	0.41	0.44	0.38	**0.50**	0.45	0.42
	GaborCNN2Speech	0.39	0.44	0.43	0.39	**0.48**	0.43	**0.43**

Figure 16 shows the accuracy of the reconstructed spectrograms for both methods across a predefined set of 51 words. Each spoke corresponds to a specific word, while the distance from the centre represents the Corr2D accuracy metric. The methods are differentiated by different colours, as shown in the legend. The results show that GaborFea2Speech achieves higher overall accuracy than GaborCNN2Speech across the word categories. Specifically, it exhibits superior performance in vocabulary such as "B", "O", and "V", while GaborCNN2Speech performs more profitably in vocabulary like "Y", "P", and "set". Furthermore, certain vocabularies such as letters (D, G, H, J, Q, U and Y), digits (1, 2, 5 and 7), colours (white and red), prepositions (in) and adverbs (please) consistently perform better in both GaborCNN2Speech and GaborFea2Speech. Conversely, words such as "F", "I", "X", "Z", "lay" and "soon", demonstrate relatively lower performance in both methods. These limitations may be attributed to the limited visual variations that occur during the pronunciation of these vocabularies, which make it difficult for speech reconstruction software to capture subtle visual changes.

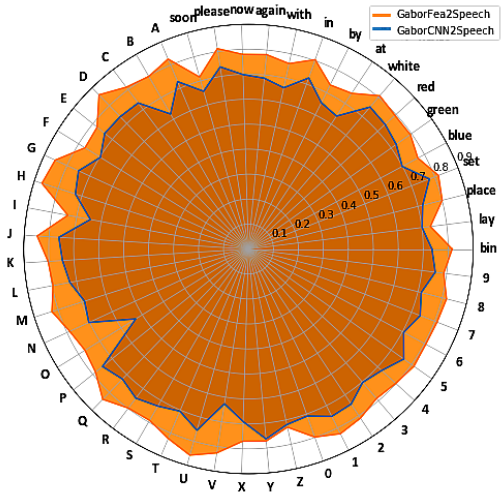


Figure 16. Radar graph analysis of GaborCNN2Speech and GaborFea2Speech in vocabulary-based speech reconstruction.

6.4. Speed and Memory Performance

One key motivation for using Gabor features is the resulting lightweight models. Fast and lightweight models are valuable in scenarios demanding quick response times and minimal resource consumption. Evaluating training time and memory usage is valuable for assessing a model's lightweight and speed attributes. These metrics directly influence the practicality of model deployment, especially in resource-constrained environments.

To compare the training time and memory usage, the GaborFea2Speech, GaborCNN2Speech, and Lip2AudSpec models were trained across various speaker scenarios. Each model was trained repeatedly in the same speaker scenario to calculate the mean and Interquartile Range (IQR) values for both training time and memory usage. The mean provides a central tendency of the data, while the IQR offers insights into the variability and stability of the model performance. This training did not include time for pre-processing, and focuses only on model training performance.

The results, as outlined in Table 8, reveal insightful contrasts in the training time and memory usage among the models, thereby allowing an evaluation of their lightweight and speed. Due to unavailability of the server used for other experiments, a different machine was used for these tests, using an Intel Xeon w3-2423 processor, 128 GB of RAM, and a NVidia RTX A5000 GPU. This improved training time for all models over the initial experiments. We calculated the mean of five runs of each configuration.

Table 8. Comparison of training time and memory usage.

Model	**1-S**		**2-S**		**4-S**		**6-S**		**8-S**	
	AVG	**IQR**	**AVG**	**IQR**	**AVG**	**IQR**	**AVG**	**IQR**	**AVG**	**IQR**
Training Time (Seconds)										
GaborFea2Speech	**6.0**	1.1	**6.4**	0.2	**9.0**	0.7	**11.4**	**0.4**	**13.6**	**0.1**
GaborCNN2Speech	139.4	**0.4**	276.5	0.2	548.4	2.8	825.4	1.4	1102.7	1.8
Lip2AudSpec	53.7	0.9	103.0	**0.1**	204.7	**0.2**	308.2	1.5	413.2	0.8
Memory Usage (MB)										
GaborFea2Speech	**2572.1**	**52.8**	**2629.1**	**23.0**	**2685.9**	**21.1**	**2739.0**	**25.5**	**2797.2**	**19.4**
GaborCNN2Speech	6042.6	319.9	7271.7	80.2	9334.6	204.1	12,047.3	551.8	14,917.3	92.3
Lip2AudSpec	6241.6	456.6	7805.7	97.0	10,236.4	153.3	13,219.4	375.2	15888.1	645.6

These results demonstrate the advanced lightweight and speed capabilities of our proposed GaborFea2Speech model, showcasing its superiority over the other models in

terms of both training efficiency and memory optimization. It can be immediately seen that the GaborFea2Speech approach uses a fraction of the memory and training time, thanks to the use of lower dimensionality features and no requirement for CNN layers. Even increasing the training set size to the full set of eight speakers did not make a notable impact. Conversely, the GaborCNN2Speech approach performed much worse than expected. Less memory was required than the CNN-based approach, but the training time was almost three times slower. An investigation identified that although less CNN layers were used, the consequence was that with less pooling and abstraction, the input into the LSTM layer became much larger, resulting in a much slower training time.

The significantly lower average training times and memory usage across various speaker scenarios underscore the GaborFea2Speech model's suitability for efficient real-time speech reconstruction, particularly in environments where computational resources are limited. One caveat is that feature extraction is slower, as while all approaches use preprocessing, the Gabor-based approach requires the application of a Gabor transform, etc. However, the results show that for model training, this is a significantly quicker and less memory intense approach overall.

7. Discussion

One key finding shows that while many results in the literature use CNN-based performance, using Gabor features can not only generate improved results, but can generate better results with a much lighter model. We were able to remove a number of layers entirely from the model without negatively affecting performance. This, along with the reduced number of input features, had a significant improvement in training time and a reduction in memory usage.

We also found that using fewer features makes for a simpler and more intuitive process, as the features can be visualised. The results also showed that by removing other information, the system was able to maintain performance with multiple speakers with less performance loss than the CNN-based model. We experimented with using the filtered Gabor image as input, but ultimately the best performance came with our proposed GaborFea2Speech approach. Multi-speaker performance is less analysed in the literature than single-speaker models, and in this paper, the strengths of our proposed approach have been demonstrated.

One interesting finding was that our Gabor-based features were consistently better than using a pure CNN-based method. It was hypothesised that this approach would deliver similar results but with a much smaller model, but the improvement with using Gabor features, particularly with multi-speaker models, was larger than initially expected. CNN-based approaches learn their own features, and previous research by the authors into deep learning with speech recognition [71] used the Lucid Visualisation Python Library [72] to investigate CNN-based features, finding that distinct features could be seen, focusing on areas such as the outline and thickness of the lips and the mouth and teeth opening (see Santos et al. for more details [71]. Arguably, by applying a Gabor transform, and then extracting features, there is less speaker-specific irrelevant information being learned, resulting in a reduction of the variability between different speakers. In terms of features, although we do extract the standard width and height of the mouth region, we also extract additional features, including orientation, area, and mass, which allows for a more three-dimensional representation to be constructed, which means that we are arguably still retrieving detailed information.

As a result of our preprocessing experiments, it was also discovered that effective speech reconstruction can be achieved by matching the Gabor sequence length with the output length of the feature and appropriately expanding the dimensionality of the feature. This shows that there are still limitations with the existing approach, with possible further improvements from better feature design rather than just using derivatives. The findings emphasise the superior performance of GaborFea2Speech in multi-speaker speech reconstruction, capitalizing on the independence of Gabor Features across speakers. How-

ever, the GaborCNN2Speech model, a GaborCNN-based approach, exhibits limitations in multi-speaker scenarios, likely due to the CNN model's difficulty in automatically distinguishing between speakers. Additionally, the performance of the Lip2AudSpec model provides further evidence of the constraints of CNN-based speech reconstruction models in multi-speaker scenarios.

In terms of the difference between approaches, there were differences between individual speakers in the single-speaker models, but these performance differences tended to be mirrored with both CNN and Gabor approaches, meaning that if the CNN approach worked well with an individual speaker, then the Gabor approaches would also report better scores. This reflects differences in articulation and speaker style and suggests that the Gabor approach is not learning anything fundamentally different from the CNN approach, just arguably a more optimised version. We have also found that both approaches, CNN and Gabor, do not perform when presented with completely novel speakers, and so there is still a fundamental limitation with regard to generalisation.

In addition, compared to other existing studies in multi-speaker speech reconstruction, such as Takashima et al.'s exemplar-based approach [26], Vougioukas et al.'s GAN-based approach [23], Um et al.'s GAN-based approach [19], and Prajwal et al.'s sequence-to-sequence system [16] adapted from Tacotron 2, which necessitate the inclusion of additional speaker characteristic features as auxiliary conditions for speaker identification, our Gabor-based model eliminates the need for such additional features, allowing for direct application in multi-speaker speech reconstruction.

Finally, it should also be noted that although our results are an improvement over CNN-based approaches, the resulting speech quality is still very limited, with significant distortion present. The work in this field is still relatively young, and there is still a lot of improvement to be found to achieve higher speech quality and improve intelligibility, especially with multiple speakers. However, fundamentally, while improved results have been reported in this paper, there will always be limitations with a purely front-facing lipreading approach, due to tongue and vocal cord movement that is not visible.

While this paper reports a number of interesting findings and demonstrates the viability of using low-dimensional Gabor features for speech reconstruction, there are a number of future research directions that will be explored. Firstly, in this paper, the focus was on comparing our proposed method using the same datasets and metrics as in other research, so future work could consider additional datasets and, in particular, data also recorded from 30 degrees and side on, rather than simply full-face cameras only. This would require additional datasets, and also additional experimentation with the Gabor feature extraction. While focusing on lightweight features, future work will also consider additional Gabor features, as well as integrating other feature extraction methods and modalities. Finally, after improving system performance, future work would require the improved system and additional feature extraction method to be tested with both objective metrics and subjective listening tests, as well as in environments with acoustic and visual noise.

8. Conclusions

In this paper, we presented a detailed literature review of audiovisual speech reconstruction and proposed two novel approaches for reconstructing speech from audio information, GaborFea2Speech and GaborCNN2Speech. These used Gabor-based feature extraction to generate relevant features. Both of these approaches outperformed CNN-based methods, and in particular, GaborFea2Speech performed particularly well with models trained with multiple speakers. Our approach used a much smaller model than is commonly used in the literature, with good results generated with a single LSTM layer. However, although good results have been demonstrated, there is still room for improvement. Future research will focus on improving output speech quality with improved models, larger datasets, and additional context information, while still prioritising a lightweight approach.

Author Contributions: Conceptualization, A.A. and Z.D.; methodology, A.A., Z.D. and Y.X.; software, Y.X. and Z.D.; writing, A.A., Z.D. and D.W.; funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by XJTLU Research Development Fund RDF-16-01-35.

Data Availability Statement: The code and data used in this study are available online at: https://github.com/zhongpingDong12/lip2Speech (accessed on 16 October 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ephrat, A.; Peleg, S. Vid2speech: Speech reconstruction from silent video. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017 ; IEEE: Piscataway, NJ, USA , 2017; pp. 5095–5099.
2. Akbari, H.; Arora, H.; Cao, L.; Mesgarani, N. Lip2audspec: Speech reconstruction from silent lip movements video. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2516–2520.
3. Abel, A.; Gao, C.; Smith, L.; Watt, R.; Hussain, A. Fast lip feature extraction using psychologically motivated gabor features. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1033–1040.
4. Munhall, K.G.; Gribble, P.; Sacco, L.; Ward, M. Temporal constraints on the McGurk effect. *Percept. Psychophys.* **1996**, *58*, 351–362. [CrossRef]
5. Le Cornu, T.; Milner, B. Generating intelligible audio speech from visual speech. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1751–1761. [CrossRef]
6. Zhang, X.; Xu, Y.; Abel, A.K.; Smith, L.S.; Watt, R.; Hussain, A.; Gao, C. Visual speech recognition with lightweight psychologically motivated gabor features. *Entropy* **2020**, *22*, 1367. [CrossRef]
7. Abel, A.; Hussain, A. *Cognitively Inspired Audiovisual Speech Filtering: Towards an Intelligent, Fuzzy Based, Multimodal, Two-Stage Speech Enhancement System*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 5.
8. Hou, J.C.; Wang, S.S.; Lai, Y.H.; Tsao, Y.; Chang, H.W.; Wang, H.M. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2018**, *2*, 117–128. [CrossRef]
9. Yu, Y.; Shandiz, A.H.; Tóth, L. Reconstructing speech from real-time articulatory MRI using neural vocoders. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 945–949.
10. Michelsanti, D.; Tan, Z.H.; Zhang, S.X.; Xu, Y.; Yu, M.; Yu, D.; Jensen, J. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1368–1396. [CrossRef]
11. Xu, Y.; Wang, H.; Dong, Z.; Li, Y.; Abel, A. Gabor-based audiovisual fusion for Mandarin Chinese speech recognition. In Proceedings of the 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 603–607.
12. Le Cornu, T.; Milner, B. Reconstructing intelligible audio speech from visual speech features. In Proceedings of the Interspeech, Dresden, Germany, 6–10 September 2015; pp. 3355–3359.
13. Aihara, R.; Masaka, K.; Takiguchi, T.; Ariki, Y. Lip-to-Speech Synthesis Using Locality-Constraint Non-Negative Matrix Factorization. In Proceedings of the MLSLP. 2015. Available online: https://www.semanticscholar.org/paper/LIP-TO-SPEECH-SYNTHESIS-USING-LOCALITY-CONSTRAINT-Aihara-Masaka/7f66836a3e822e7677f11350bf170d09f6150b9f (accessed on 16 October 2023).
14. Ra, R.; Aihara, R.; Takiguchi, T.; Ariki, Y. Visual-to-speech conversion based on maximum likelihood estimation. In Proceedings of the 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), Nagoya, Japan, 8–12 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 518–521.
15. Abel, A.; Hussain, A. Novel Two-Stage Audiovisual Speech Filtering in Noisy Environments. *Cogn. Comput.* **2014**, *6*, 200–217. [CrossRef]
16. Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V.P.; Jawahar, C. Learning individual speaking styles for accurate lip to speech synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13796–13805.
17. Kim, M.; Hong, J.; Ro, Y.M. Lip to speech synthesis with visual context attentional GAN. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2758–2770.
18. Oneață, D.; Stan, A.; Cucu, H. Speaker disentanglement in video-to-speech conversion. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 46–50.
19. Um, S.Y.; Kim, J.; Lee, J.; Kang, H.G. Facetron: A Multi-speaker Face-to-Speech Model based on Cross-modal Latent Representations. *arXiv* **2021**, arXiv:2107.12003.

20. Wang, D.; Yang, S.; Su, D.; Liu, X.; Yu, D.; Meng, H. VCVTS: Multi-speaker video-to-speech synthesis via cross-modal knowledge transfer from voice conversion. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 7–13 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 7252–7256.
21. Ephrat, A.; Halperin, T.; Peleg, S. Improved speech reconstruction from silent video. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 455–462.
22. Perraudin, N.; Balazs, P.; Søndergaard, P.L. A fast Griffin-Lim algorithm. In Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1–4.
23. Vougioukas, K.; Ma, P.; Petridis, S.; Pantic, M. Video-driven speech reconstruction using generative adversarial networks. *arXiv* **2019**, arXiv:1906.06301.
24. Mira, R.; Vougioukas, K.; Ma, P.; Petridis, S.; Schuller, B.W.; Pantic, M. End-to-end video-to-speech synthesis using generative adversarial networks. *IEEE Trans. Cybern.* **2022**, *53*, 3454–3466 . [CrossRef] [PubMed]
25. Yadav, R.; Sardana, A.; Namboodiri, V.P.; Hegde, R.M. Speech prediction in silent videos using variational autoencoders. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 7048–7052.
26. Takashima, Y.; Takiguchi, T.; Ariki, Y. Exemplar-based lip-to-speech synthesis using convolutional neural networks. In Proceedings of the IW-FCV. 2019. Available online: https://www.semanticscholar.org/paper/Exemplar-based-Lip-to-Speech-Synthesis-Using-Neural-Takashima-Takiguchi/cbad0d803fdbdceacd112093b573ac70b6ccd146 (accessed on 16 October 2023).
27. Hong, J.; Kim, M.; Park, S.J.; Ro, Y.M. Speech reconstruction with reminiscent sound via visual voice memory. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3654–3667. [CrossRef]
28. Mira, R.; Haliassos, A.; Petridis, S.; Schuller, B.W.; Pantic, M. SVTS: Scalable video-to-speech synthesis. *arXiv* **2022**, arXiv:2205.02058.
29. He, J.; Zhao, Z.; Ren, Y.; Liu, J.; Huai, B.; Yuan, N. Flow-based unconstrained lip to speech generation. *Proc. Aaai Conf. Artif. Intell.* **2022**, *36*, 843–851. [CrossRef]
30. Varshney, M.; Yadav, R.; Namboodiri, V.P.; Hegde, R.M. Learning Speaker-specific Lip-to-Speech Generation. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 491–498.
31. Millerdurai, C.; Khaliq, L.A.; Ulrich, T. Show Me Your Face, and I'll Tell You How You Speak. *arXiv* **2022**, arXiv:2206.14009.
32. Wang, Y.; Zhao, Z. Fastlts: Non-autoregressive end-to-end unconstrained lip-to-speech synthesis. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 5678–5687.
33. Hegde, S.B.; Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V.P.; Jawahar, C. Lip-to-speech synthesis for arbitrary speakers in the wild. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 6250–6258.
34. Qu, L.; Weber, C.; Wermter, S. LipSound: Neural Mel-Spectrogram Reconstruction for Lip Reading. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September ; 2019; pp. 2768–2772.
35. Qu, L.; Weber, C.; Wermter, S. LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading. *IEEE Trans. Neural Networks Learn. Syst.* **2022**. [CrossRef]
36. Zeng, R.; Xiong, S. Lip to Speech Synthesis Based on Speaker Characteristics Feature Fusion. In Proceedings of the 4th International Conference on Information Technology and Computer Communications, Guangzhou, China, 23–25 June 2022; pp. 78–83.
37. Kim, M.; Hong, J.; Ro, Y.M. Lip-to-speech synthesis in the wild with multi-task learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
38. Kumar, Y.; Jain, R.; Salik, M.; ratn Shah, R.; Zimmermann, R.; Yin, Y. Mylipper: A personalized system for speech reconstruction using multi-view visual feeds. In Proceedings of the 2018 IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, 10–12 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 159–166.
39. Kumar, Y.; Aggarwal, M.; Nawal, P.; Satoh, S.; Shah, R.R.; Zimmermann, R. Harnessing ai for speech reconstruction using multi-view silent video feed. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1976–1983.
40. Salik, K.M.; Aggarwal, S.; Kumar, Y.; Shah, R.R.; Jain, R.; Zimmermann, R. Lipper: Speaker independent speech synthesis using multi-view lipreading. *Proc. Aaai Conf. Artif. Intell.* **2019**, *33*, 10023–10024. [CrossRef]
41. Kumar, Y.; Jain, R.; Salik, K.M.; Shah, R.R.; Yin, Y.; Zimmermann, R. Lipper: Synthesizing thy speech using multi-view lipreading. *Proc. Aaai Conf. Artif. Intell.* **2019**, *33*, 2588–2595. [CrossRef]
42. Uttam, S.; Kumar, Y.; Sahrawat, D.; Aggarwal, M.; Shah, R.R.; Mahata, D.; Stent, A. Hush-Hush Speak: Speech Reconstruction Using Silent Videos. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 136–140.
43. Raheja, J.L.; Kumar, S.; Chaudhary, A. Fabric defect detection based on GLCM and Gabor filter: A comparison. *Optik* **2013**, *124*, 6469–6474. [CrossRef]
44. Dakin, S.C.; Watt, R.J. Biological "bar codes" in human faces. *J. Vis.* **2009**, *9*, 2. [CrossRef] [PubMed]

45. Martinez, A.M.C.; Mallidi, S.H.; Meyer, B.T. On the relevance of auditory-based Gabor features for deep learning in robust speech recognition. *Comput. Speech Lang.* **2017**, *45*, 21–38. [CrossRef]
46. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. In Proceedings of the Computer Vision—ECCV'98: 5th European Conference on Computer Vision, Freiburg, Germany, 2–6 June 1998; Proceedings, Volume II; Springer: Berlin/Heidelberg, Germany, 1998; pp. 484–498.
47. Ahmed, N.; Natarajan, T.; Rao, K.R. Discrete cosine transform. *IEEE Trans. Comput.* **1974**, *100*, 90–93. [CrossRef]
48. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
49. Xu, Y.; Li, Y.; Abel, A. Gabor Based Lipreading with a New Audiovisual Mandarin Corpus. In *Advances in Brain Inspired Cognitive Systems*; Ren, J., Hussain, A., Zhao, H., Huang, K., Zheng, J., Cai, J., Chen, R., Xiao, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 169–179.
50. King, D.E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
51. Jones, J.P.; Palmer, L.A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **1987**, *58*, 1233–1258. [CrossRef]
52. Chowdhury, M.H.; Little, W.D. Image thresholding techniques. In Proceedings of the IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing, Victoria, BC, Canada, 17–19 May 1995; IEEE: Piscataway, NJ, USA, 1995; pp. 585–589.
53. Yen, J.Y. Finding the k shortest loopless paths in a network. *Manag. Sci.* **1971**, *17*, 712–716. [CrossRef]
54. Yadav, K.; Singh, A.; others. Comparative analysis of visual recognition capabilities of CNN architecture enhanced with Gabor filter. In Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2–4 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 45–50.
55. Chang, S.Y.; Morgan, N. Robust CNN-based speech recognition with Gabor filter kernels. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
56. Zadeh, M.M.T.; Imani, M.; Majidi, B. Fast facial emotion recognition using convolutional neural networks and Gabor filters. In Proceedings of the 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI), Tehran, Iran, 28 February–1 March 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 577–581.
57. Chi, T.; Ru, P.; Shamma, S.A. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* **2005**, *118*, 887–906. [CrossRef]
58. Cornu, T.L.; Milner, B. Reconstructing intelligible audio speech from visual speech features. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
59. Sindhura, P.; Preethi, S.; Niranjana, K.B. Convolutional neural networks for predicting words: A lip-reading system. In Proceedings of the 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Msyuru, India, 14–15 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 929–933.
60. Cooke, M.; Barker, J.; Cunningham, S.; Shao, X. An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **2006**, *120*, 2421–2424. [CrossRef] [PubMed]
61. Ketkar, N.; Ketkar, N. Introduction to keras. In *Deep Learning with Python: A Hands-on Introduction* ; Apress: Berkeley, CA, USA, 2017; pp. 97–111.
62. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
63. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
64. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
65. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* **2015**, arXiv:1511.07289.
66. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
67. Barlow, A.L.; MacLeod, A.; Noppen, S.; Sanderson, J.; Guérin, C.J. Colocalization analysis in fluorescence micrographs: Verification of a more accurate calculation of pearson's correlation coefficient. *Microsc. Microanal.* **2010**, *16*, 710–724. [CrossRef] [PubMed]
68. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 2, pp. 749–752.
69. Hu, Y.; Loizou, P.C. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *16*, 229–238. [CrossRef]
70. Kwon, J.K.; Park, S.; Sung, D.K. Log-likelihood ratio (LLR) conversion schemes in orthogonal code hopping multiplexing. *IEEE Commun. Lett.* **2003**, *7*, 104–106. [CrossRef]

71. Israel Santos, T.; Abel, A.; Wilson, N.; Xu, Y. Speaker-Independent Visual Speech Recognition with the Inception V3 Model. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 613–620. [CrossRef]
72. Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; Mordvintsev, A. The building blocks of interpretability. *Distill* **2018**, *3*, e10. [CrossRef]