Gabor-based Audiovisual Fusion for Mandarin Chinese Speech Recognition

Yan Xu, Hongce Wang, Zhongping Dong, and Yuexuan Li
School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, China
Yan.Xu@student.xjtlu.edu.cn
{hongce.wang19, zhongping.dong12}@alumni.xjtlu.edu.cn
effyli1997@gmail.com

Andrew Abel
Faculty Of Science
University Of Strathclyde
Scotland, UK
andrew.abel@strath.ac.uk

***Abstract*—Audiovisual Speech Recognition (AVSR) is a popular research topic, and incorporating visual features into speech recognition systems has been found to deliver good results. In recent years, end-to-end Convolutional Neural Network (CNN) based deep learning has been widely utilized. However, these often require big data and can be time consuming to train. A lot of speech research also tends to focus on English language datasets. In this paper, we propose a lightweight optimized and automated speech recognition system using Gabor based feature extraction, combined with our Audiovisual Mandarin Chinese (AVMC) corpus. This combines Mel-frequency Cepstral Coefficients (MFCCs) + CNN_Bidirectional Long Short-term Memory (BiLSTM)_Attention (CLA) model for Audio Speech Recognition, and low dimension Gabor visual features + CLA model for Visual Speech Recognition. As we are focusing on Chinese language recognition, we individually analyse initials, finals, and tones, as part of pinyin speech production. The proposed low dimensionality system achieves 88.6%, 87.5% and 93.6% accuracy for tones, initials and finals respectively at char-level, 84.8% for pinyin at word-level.**

***Index Terms*—Audiovisual, Speech Recognition, Chinese, Gabor Filter**

I. Introduction

Audiovisual Speech Recognition (AVSR) is regarded as a robust speech recognition method. Many recent AVSR systems [1]–[5] show improved performance on large English language datasets. However, research using Chinese datasets is much more limited [6]. For visual speech recognition specifically, a number of recently developed neural network based methods can produce state-of-the-art performance on both English and Chinese datasets, including Assael et al. [7], Chung and Zisserman [8] used English datasets, and Zhao et al. [9], Ma et al. [10] [11] and Yuan et al. [12] focused on a Chinese dataset. These used convolutional layers to extract lip features, and fed them into a recurrent neural network, an attention architecture [8], or transformer architecture [10]. However, these networks produce autoencoded features that are non-intuitive for humans to understand, and have high computational cost and dataset requirements. The Chinese language recognition models also generally consider pinyin as a whole, without focusing on the individual language components (initials, finals, and tones).

Another option is to consider more 'classical' methods, which follow two steps: feature extraction and feature recognition. For extraction methods such as Active Appearance Models (AAM), Discrete Cosine Transformation (DCT), and Gabor filters. Saudi et al. [13] summarised the benefits of Gabor filters: 1. they are insensitive to different patterns in the spectro-temporal representation of the visual signal, 2. they minimize the product of their standard deviation in both frequency and time domain, 3. they have been successfully used in different applications [14] [15]. Horizontal Gabor features are arguably a match to psychological models of human face recognition, and have been found to be a reliable visual feature extraction method [16]. Previous research [15] proposed a handcrafted Gabor-based VSR system with good initial results. For recognition, many approaches have been proposed, with LSTM networks [17] and the bidirectional LSTM (BiLSTM) [18] widely used. Another architecture, Inception-ResNet, which combines the Inception CNN module with residual connections, was also successfully used in recent Chinese lipreading work [15]. Recently, attention mechanisms have also been employed for focusing on interactive information from the key frames in temporal sequences [8].

The CNN_BiLSTM_Attention (CLA) model [19] has been used for time series prediction, and is suited to our research needs. Therefore, in this paper, a lightweight AVSR system is proposed. For VSR, an automated Gabor based lip feature extraction system with the CLA model has been used for fast lip feature extraction and learning the hidden connections in spatiotemporal information, with attention weights introduced for expressing the importance of key frames. For Automatic Speech Recognition (ASR), Mel-frequency Cepstral Coefficient (MFCC) features and the CLA model are utilized to extract and recognize audio features.

There are several key contributions of this paper. Firstly, an improved fully automated Gabor-based feature extraction system is proposed, using Bayesian optimization to identify Gabor hyper parameters, with good results. We also demonstrate that we can improve on these by implementing decision (late) fusion to produce better results than a single modality. Our Gabor-based features, despite being lightweight, and with

a lower dimensionality than CNN features, and far quicker to train, can deliver results equivalent to using a full CNN autoencoder approach, meaning that applying Gabor features for AVSR is effective. We also separately analyse performance for Chinese pinyin, initials, finals, and tones to investigate detailed lipreading performance.

II. Visual Speech Recognition

Geometric features could arguably be seen as a less "fashionable" approach than using CNN based inputs, and yet, a Gabor wavelet transform is a fast and lightweight approach for extracting detailed mouth region measurements without the need for any detailed mouth modelling [16]. We have previously demonstrated a fast and reliable handcrafted system which delivers good results [15]. However, the Gabor parameters needed to be manually defined to extract precise geometric features. This is a general limitation with the 'handcrafted' approach in comparison with deep learning models, the requirement for hyper parameters to be manually tuned. To solve this problem, we have improved the initial model and added an optimization algorithm to automatically extract features. Here, an automated and optimized Gabor-based feature extraction and recognition model has been proposed.

Due to space limitations, more details about system design can be found in Xu et al. [15], but briefly, there are several key lip feature extraction steps:

1. Extract individual frame from a video.
2. Extract ROI (mouth region) using the Dlib toolkit.
3. Obtain an optimised group of Gabor parameters using Bayesian optimization to tune hyper parameters. The Gabor kernel is packaged as Opencv function, $cv2.getGaborKernel((ksize, ksize), \sigma, \theta, \lambda, \gamma, \phi)$, the real part of Gabor transform is performed as follows:

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

where:

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

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

where $ksize$ is the size of Gabor kernel, 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 (set to 90 degrees). λ represents the wavelength of the sinusoidal factor, γ is the spatial aspect ratio, and ψ is the phase offset, defined as 0.

Algorithm 1 Optimized system with Gabor filtering

Require: Original width and height: $Width_D, Hight_D$
Input: Gabor parameters sets (search space): $Ksize, \lambda, \gamma, \sigma$. Optimized Gabor parameters: $Ksize_i, \lambda_i, \gamma_i, \sigma_i$

```
for Ksize_i, λ_i, γ_i, σ_i in search speace do
    Extract Gabor features use optimized parameters
    Select mouth part according to Width_D, Hight_D
    Evaluate f = |Width_D − Width_G| + |Height_D − Height_G|
    if f is minimal then
        return Ksize_i, λ_i, γ_i, σ_i
    end if
end for
```

The optimization method is described in Algorithm 1. Before filtering the ROI, optimal parameters need to be found, therefore, there is a loop to cycle the value of each parameter with Tree-structure Parzen Estimator Approach (TPE) based Bayesian optimization, a powerful and very widely used optimization tool proposed by Bergstra et al. [21], packaged as the 'hyperopt' Python function. The orientation θ is set as 90, since we only extract horizontal features. The Phase offset ϕ is set as 0 by default, the other four parameters $(Ksize, \lambda, \gamma, \sigma)$ need to be tuned. Based on preliminary experimentation, the search space is set as: $ksize$ ranges from 5 to 25, λ ranges between 5 and 25, σ can be from 2 to 10, γ is between 0.2 and 0.9. A set of Gabor parameters $Ksize_i, \lambda_i, \gamma_i, \sigma_i$ are optimized by TPE algorithm within this search space. Those optimized parameters are used to estimate the error $f = |Width_D - Width_G| + |Height_D - Height_G|$. $Width_D$ and $Height_D$ are denoted as the lip width and height extracted by the Dlib toolkit. $Width_G$ and $Height_G$ are denoted as the lip width and height extracted by our method. Hyper parameters that minimize the error are used for Gabor filtering.

4. Apply Gabor transform to obtain mouth region features with optimized hyper parameters.
5. Select the mouth region and return 7 parameters: width, height, area, intensity, x and y values of central point, and orientation (more details in Xu et al [15]). The 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 1. As well as delivering good performance, we use similar features (albeit without automation) in previous work [20], and show that in noisy environments, we can identify distinct changes in visual speech patterns. Preliminary work [16] also identified consistent visualisation patterns with different speakers, so that words could be visualised.

The model applied for VSR is the same for four recognition targets: tones, initials, finals, and Pinyin. Each uses a coordinated CLA model. We input the Gabor feature vectors for a single character, and a dense layer is used for classification, with a different layer used, depending on whether we wish to classify initials, finals, Pinyin, or tones. This method uses a single convolutional layer to provide high quality features to the BiLSTM layer, and the attention model learns key features. The structures of both audio-only and visual-only models are very similar. As we use Gabor features rather than CNN inputs, the CNN+CLA model is not transfer learning based.

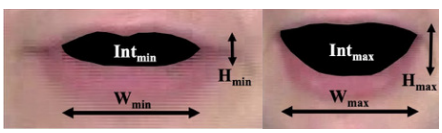

Fig. 1: Example of mouth ROI, showing height, width, area, and intensity, taken from Chiu et al. [20].

III. Multimodal Fusion for Speech Recognition.

Audiovisual fusion refers to fusing the features from different speech modalities (i.e. audio and visual) to improve speech recognition. In this paper, we use the visual features as discussed above, and extract MFCCs as audio features, where maximum frame length (with zero padding) is set to 50 frames, with each frame having 57 MFCC coefficients.

There are two common strategies, early or late fusion [22]. Early fusion combines (e.g. concatenates) the features from both modalities before machine learning, whereas late (decision) fusion applies machine learning separately to each modality stream, before combining them to produce a final result. Research has shown that both strategies can have good results, depending on the scenario. Gogate et al. [23] recently identified good results with late fusion, and we adopted this strategy here. Our fusion model combines the ASR and VSR models, and adds two dense layers for final classification.

To present the detailed structure, we choose the tone fusion model as an example. For the VSR model, the layers and output size are shown in Table I (left), the input is the sequence of Gabor features, the first step is to normalize inputs, then feed it into a 1D CNN to learn time series features from the Gabor features. The features are transferred to the pooling layer for further aggregation. After convolution, we use BiLSTM to model the hidden relationship of frames in the character.The dropout layer is then used to prevent overfitting. For the attention mechanism, the axes of the input layer with index 1 and 2 are permuted and we add a dense layer with 'softmax' activation function to select a vector from the input that contributes most to the target. The weight matrix is identified by permuting the axes again, and the attention model layer and BiLSTM layer are multiplied, with a dense layer to make a combined decision for identifying the target. The ASR model is similar to the VSR model (Table I (right)). The two outputs are combined by the concatenation layer, then with the final two dense layers for classifying the target (Table I).

TABLE I: Fusion model of tones.

AVSR			
VSR		ASR	
InputLayer (None, 35, 7)		**InputLayer** (None, 50, 57)	
Batch Normalization		**Batch Normalization**	
Conv1D filters: 64 kernel size: 1 activation: relu		**Conv1D** filters: 64 kernel size: 1 activation: relu	
MaxPool1D pool size: 2 strides: 2		**Dropout** rate: 0.3	
BLSTM filter: 128		**BLSTM** filter: 128	
Dropout rate: 0.3 Output shape (None,17, 256)		**Dropout** rate: 0.3 Output shape (None, 50, 128)	
Permute(2,1)		**Permute**(2,1)	
Dense filter: 17 activation: softmax		**Dense** filter: 50 activation: softmax	
Permute(2,1)		**Permute**(2,1)	
Multiply		**Multiply**	
MaxPool1D pool size: 14 strides: 4		**MaxPool1D** pool size: 50 strides: 2	
Dense filter: 100 actication: sigmoid		**Dense** filter: 200 actication: sigmoid	
textConcatenate			
Dense filter: 64			
Dropout rate: 0.25			
Dense filter: 4			
OutputLayer (None, 1, 4)			

IV. Mandarin Chinese Corpus

Many English multimodal corpora have been published such as Grid [24], and LRS [8]. Chinese differs from English in that it consists of characters and Pinyin. Each Chinese Character possesses one syllable which is represented by pinyin. Each pinyin has an initial, a final, and a tone. For example, the pinyin 'mén', where 'm' is the initial, and 'én' is the final, with a tone mark of 2. In addition, Chinese characters are monosyllabic [25], differing from languages like English. However, there are only a small number of existing AVSR Chinese corpus such as LRW-1000 [26], the CMLR Dataset [9], and the CCTV website based dataset [27]. These are all large scale with a complex visual background and a noisy speech environment. For this project, to perform accurate Pinyin character recognition, we require a labelled video corpus of distinct Chinese characters, recorded in a clean environment. Therefore, the AVMC dataset [15], developed in our previous work, is more suitable for our research needs. More details can be found in Xu et al. [15].

Some previous research [27] [9] has attempted to predict pinyin through single letters. For example, the pinyin of 'men' will be predicted as ['m', 'e', 'n']. However, this prediction method can easily confuse the recognition system because some letters appear in both initials and finals with different pronunciation. An example of this is the letter 'g', which appears in 'guo' as an initial, and also appears in 'jing' as a component of a final. Ma et al. [10] argued that the pronunciation of pinyin is a syllable, they therefore predict pinyin as units of initials and finals. For example, the pinyin 'men' will be predicted as an initial 'm' and finals 'en'. Furthermore, some finals are compound finals, consisting of two or three simple finals, such as 'uan' and 'ian', the pronunciation of the last two finals 'a' and 'n' are the same, but the first final 'u' and 'i' are different. We propose to overcome this and remedy the defects in the previous approaches by dividing pinyin into initials and finals and predicting letters separately.

V. Results and Discussion

A. System Configuration

All machine learning and feature extraction took place on a desktop machine, with Windows 10 Pro installed, using the

TABLE II: Accuracy and IQR of different VSR models.

VSR model	Tone	Initial	Final	Pinyin
Gabor + Inception [15]	0.561	0.628	0.730	0.617
New Gabor + CLA	0.485	0.568	0.762	0.642
CNN +CLA (100)	0.487	0.554	0.622	0.533
CNN +CLA (300)	0.509	0.558	0.747	0.587

Interquartile Range	Tone	Initial	Final	Pinyin
New Gabor +CLA	0.010	0.010	0.010	0.016
CNN +CLA (300)	0.034	0.011	0.020	0.032

Anaconda 3 Jupyter notebook. The CPU was an Intel i7-8700K with a 3.70 GHz clock speed, and 32 GB of RAM.

ASR and VSR are both similar models, the configuration of initial, final and pinyin recognition models are similar to the tone model which is shown in Table I. The ratio of training data + validation data to testing data is 8:2. The loss in recognizing the digits is evaluated using the categorical_crossentropy loss function and for optimization we have applied Adam optimization to minimize the error. All our experiments were run for 300 epochs utilizing the ReduceLROnPlateau schedule with patience of 6 epochs, decay factor of 0.5 , verbose of 1 and minimal learning rate of 0.00001 with a batch size of 128.

B. Visual Only Speech Recognition

To evaluate our proposed automated Gabor optimization and audiovisual fusion approach, we compare our results with previous results reported in Xu et al. [15], and also by using a CNN based autoencoder to create an end-to-end system [28]. The results are shown in Table II. All experiments were run 5 times with the datasets randomly shuffled, and the mean and interquartile range (IQR) calculated. The only exception was the CNN+CLA(100) configuration.

Table II shows that our proposed approach (New Gabor + CLA) delivers good results. When considering only the initials (e.g. 'zh' in 'zhang') the accuracy is 0.568, due to there being less visual information available at the start of a word, and there being more variation than for finals. Identifying the tones (Mandarin Chinese has 4 tones) has an accuracy of 0.485, however, the confusion matrix (not shown here due to space limitations) showed that the results are effectively randomised, which is to be expected, given the tone is entirely vocal. However, for finals and pinyin, we see much better results.

Compared to work in the literature [15], ('Gabor + Inception' in Table II), which used handcrafted Gabor features and the Inception model, our results are similar, with poorer performance with initials and tones (as discussed, tone results with VSR should be treated with caution), but slightly better performance with finals and overall pinyin results. While the performance is similar, we are using a different and more robust feature extraction method, and when we disregard tones (which are not detected accurately with any VSR only approach), this means that on two of the 3 outputs, the final and pinyin, our upgraded model is better performing.

For comparison, we also conducted end-to-end training using a CNN [28], with the full image as the input. This is much larger than our 7 features, and 100 epochs was not enough to complete training, whereas it was enough for our proposed features. Therefore only a single run for the 100 epoch configuration was reported, and we instead used 300 epochs. The results in Table II show broadly similar accuracy as our features, however the training time was considerably slower, as shown in Table III. Training our network to recognise pinyin using Gabor features took 241.9 seconds, however to achieve equivalent results with CNN input took 9126.6 seconds. It is also worth mentioning that to improve performance for end-to-end learning, additional layers would be required to optimise training, resulting in time increases.

TABLE III: Mean training time of VSR models in seconds.

VSR model	Tone	Initial	Final	Pinyin
New Gabor +CLA	240.574	198.079	125.375	241.903
CNN +CLA (100)	710.483	1561.795	563.196	3089.428
CNN +CLA (300)	1884.761	4781.382	4313.816	9126.552

TABLE IV: Average accuracy and IQR for ASR (MFCC features), VSR (Gabor features) and AVSR (MFCC and Gabor) models.

Models	Tone	Initial	Final	Pinyin
(ASR) MFCC + CLA	0.868	0.866	0.913	0.809
(VSR) New Gabor + CLA	0.485	0.568	0.762	0.642
(AVSR) MFCC + New Gabor + CLA	0.886	0.875	0.936	0.848

Interquartile Range	Tone	Initial	Final	Pinyin
(ASR) MFCC + CLA	0.019	0.022	0.015	0.139
(VSR) New Gabor +CLA	0.010	0.010	0.010	0.016
(AVSR) MFCC + New Gabor + CLA	0.011	0.018	0.010	0.024

C. Multimodal Fusion for Speech Recognition

Table IV shows the results of ASR, VSR and AVSR, calculated for initials, finals, pinyin, and tones. As expected, audio only outperforms visual only, and fusion improves the results. For tones and initials, we find small improvements of 0.018 and 0.009, but for finals (0.913 with audio only), fusion resulted in an accuracy of 0.936, an improvement of 0.023. For pinyin, the improvement was even larger, with fusion having an accuracy of 0.848, an improvement of 0.039.

Direct comparison with other research is challenging. For VSR, Zhang et al. [27] use the CCTV news dataset. Their LipCH-Net model achieved 58.7% (Pinyin-level) accuracy, they also test their corpus with LipNet [7], which reported 95.2% accuracy in sentence-level on GRID, but only achieved 41.6% (Pinyin-level) accuracy. Zhao et al. [9] proposed the CSSMCM model, which uses a two-layer bi-directional GRU for the encoder and a two-layer uni-direction GRU for the decoder. It achieves 63.78%, 89.05% and 67.52% on pinyin, tones and character accuracy with the CMLR dataset. Ma et al. [10] use the same dataset to predict pinyin as the units of initials and finals, with 74.64% accuracy by using the CTCH-LipNet model. Our proposed optimised lightweight approach achieves a competitive recognition rate on finals and pinyin with 76.2% and 64.2%, using the AVMC dataset.

For AVSR, Yuan et al. [12] report state-of-the-art performance on the LRW-1000 [29] Chinese dataset. The word sequence for 3D-CNN_BiLSTM_Attention for VSR and MFCC was fed into CNN_BiLSTM_Attention for ASR. By using late

fusion, it achieved 82.78% accuracy and found a 7.9% percent improvement in the audio only results when they added visual information, similar to the gains reported here, thus verifying our findings using our lightweight approach.

VI. Conclusion

In this paper, a lightweight audiovisual speech recognition system, including MFCC feature extraction, optimized automated Gabor feature extraction and recognition, is proposed for Chinese language speech recognition. Despite using a much lighter Gabor-based method rather than a big-data approach, our results show a competitive recognition rate and a very quick training time. Using geometric features is arguably a more traditional approach, and yet, we have developed a fast and reliable system. Furthermore, it is often argued that the key limitation with 'handcrafted' approaches in comparison with deep learning models is that the hyper parameters need to be manually tuned. Our approach is fully automated, thus negating a major issue with this approach.

We also identify performance differences specific to Chinese spoken language, with visual information more useful in finals and pinyin, but arguably less so in initials and tones. In addition, finals perform better than pinyin in all VSR and AVSR experiments which verified our analysis that letters that appeared in both initials and finals confused the system, and decreased the pinyin recognition rate. Finally, the late fusion of audio and visual streams is used to deliver optimised results. Future work will investigate the exact errors that visual information solves/introduces upon fusion, and extend this work to make it more robust and reliable in the real world.

References

[1] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7613–7617.

[2] W. Yu, S. Zeiler, and D. Kolossa, "Large-vocabulary audio-visual speech recognition in noisy environments," *arXiv preprint arXiv:2109.04894*, 2021.

[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[4] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 905–912.

[5] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 513–520.

[6] L. Liangi, Y. Luo, F. Huang, and A. V. Nefian, "A multi-stream audio-video large-vocabulary mandarin chinese speech database," in *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, vol. 3. IEEE, 2004, pp. 1787–1790.

[7] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

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

[9] Y. Zhao, R. Xu, and M. Song, "A cascade sequence-to-sequence model for chinese mandarin lip reading," *CoRR*, vol. abs/1908.04917, 2019. [Online]. Available: http://arxiv.org/abs/1908.04917

[10] S. Ma, S. Wang, and X. Lin, "A transformer-based model for sentence-level chinese mandarin lipreading," in *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2020, pp. 78–81.

[11] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2857–2866.

[12] Y. Yuan, W. Tang, M. Fan, Y. Cao, P. Zhang, and L. Xie, "Deep audio-visual system for closed-set word-level speech recognition," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 540–545.

[13] A. S. Saudi, M. I. Khalil, and H. M. Abbas, "Improved features and dynamic stream weight adaption for robust audio-visual speech recognition framework," *Digital Signal Processing*, vol. 89, pp. 17–29, 2019.

[14] X. Zhang, Y. Xu, A. K. Abel, L. S. Smith, R. Watt, A. Hussain, and C. Gao, "Visual speech recognition with lightweight psychologically motivated gabor features," *Entropy*, vol. 22, no. 12, p. 1367, 2020.

[15] Y. Xu, Y. Li, and A. Abel, "Gabor based lipreading with a new audiovisual mandarin corpus," in *International Conference on Brain Inspired Cognitive Systems*. Springer, 2019, pp. 169–179.

[16] A. Abel, C. Gao, L. Smith, R. Watt, and A. Hussain, "Fast lip feature extraction using psychologically motivated gabor features," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2018, pp. 1033–1040.

[17] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with lstms," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2592–2596.

[18] Y. Lu and J. Yan, "Automatic lip reading using convolution neural network and bidirectional long short-term memory," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, no. 01, p. 2054003, 2020.

[19] PatientEz. (2020) Cnn bilstm attention time series prediction keras. https://github.com/PatientEz/CNN-BiLSTM-Attention-Time-Series-Prediction_Keras/blob/master/Main.py.

[20] W. Chiu, Y. Xu, A. Abel, C. Lin, and Z. Tu, "Investigating the Visual Lombard Effect with Gabor Based Features," in *Proc. Interspeech 2020*, 2020, pp. 4606–4610. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1291

[21] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in neural information processing systems*, 2011, pp. 2546–2554.

[22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[23] M. Gogate, A. Adeel, and A. Hussain, "Deep learning driven multimodal fusion for automated deception detection," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–6.

[24] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[25] J. M. Howie and J. M. Howie, *Acoustical studies of Mandarin vowels and tones*. Cambridge University Press, 1976, vol. 18.

[26] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[27] X. Zhang, H. Gong, X. Dai, F. Yang, N. Liu, and M. Liu, "Understanding pictograph with facial features: End-to-end sentence-level lip reading of chinese," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 9211–9218, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4956

[28] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2516–2520.

[29] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," *CoRR*, vol. abs/1810.06990, 2018. [Online]. Available: http://arxiv.org/abs/1810.06990