Lip Forgery Video Detection via Multi-phoneme Selection

Jiaying Lin[1], Wenbo Zhou[1]*, Honggu Liu[1], Hang Zhou[2], Weiming Zhang[1]* and Nenghai Yu[1]

[1]*University of Science and Technology of China*
[2]*Simon Fraser University*

Abstract

Deepfake technique can produce realistic manipulation videos including full-face synthesis and local region forgery. General methods work well in detecting the former but are usually intractable in capturing local artifacts especially for lip forgery detection. In this paper, we focus on the lip forgery detection task. We first establish a robust mapping from audio to lip shapes. Then we classify the lip shapes of each video frame according to different spoken phonemes, enable the network in capturing the dissonances between lip shapes and phonemes in fake videos, increasing the interpretability. Each lip shape-phoneme set is used to train a sub-model, those with better discrimination will be selected to obtain an ensemble classification model. Extensive experimental results demonstrate that our method outperforms the most state-of-the-art methods on both the public DFDC dataset and a self-organized lip forgery dataset.

Keywords

Lip Forgery, Deepfake Detection, Phoneme and Viseme

1. Introduction

Thanks to the tremendous success of deep generative models, face forgery becomes an emerging research topic in very recent years and various methods have been proposed [1, 2]. Depending on the manipulated region, they can be roughly categorized into two types: full-face synthesis [3, 4] that usually swaps the whole synthesized source face to a target face, and local face region forgery [5, 6] that only modifies partial face region, *e.g.*, modifying the lip shape to match the audio content. Especially when the lips of politicians have been tampered with to make inappropriate speeches, it can lead to serious political crisis.

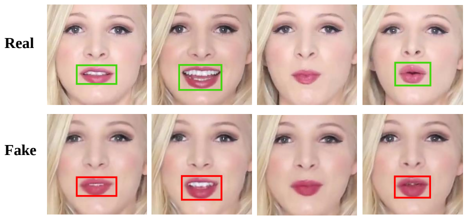


Figure 1: The lip shapes of speaking the word "apple" in real (top) and fake (bottom) video. In the real video, the lips are more widely opened with clear teeth texture, while opposite in the fake.

To alleviate the risks brought by malicious uses of face forgery, many detection methods have been proposed [7, 8, 9]. These methods usually consider the forgery detection from different aspects and extract visual features from the whole face region, achieving impressive detection results on public datasets FF++ and DFDC, in which most of the fake videos are tampered in a full-face synthesized manner. But this type of detection methods struggle to handle the local region forgery cases like lip-sync [5]. Recently, [10] attempt to detect lip-sync forgery video with single phoneme-viseme matching for specific targets. [11, 12] employ features such as audio and expression to detect synchronization between different modalities.

To address the problem of local region forgery detection, in this paper, we proposed a complete multi-phoneme selection-based framework. To take full advantage of the particularity of lip forgery videos that contain audios, we need to establish a robust mapping relationship between the lip shapes and the audio contents. Prior studies in the realm of Audio-Visual Speech Recognition have demonstrated that the phoneme is the smallest identifiable unit correlated with a particular lip shape. Motivated by [13], we divide audio contents into 12 phoneme classes and classify all the video frames. For each phoneme-lip set, we measure the deviation on open-close amplitude between real and fake lip shapes, and train a sub-model for real/fake classification.

Usually, a large deviation represents the obvious discrepancy between the real and fake lip shapes, which also indicates the great difficulty in synthesizing the lip shape for the corresponding phoneme. Simultaneously, it shows the robustness of correlated phoneme-lip mapping

Woodstock'21: Symposium on the irreproducible science, June 07–11, 2021, Woodstock, NY

* *Corresponding Author.*
vivian19@mail.ustc.edu.cn (J. Lin); welbeckz@ustc.edu.cn (W. Zhou); lhg9754@mail.ustc.edu.cn (H. Liu); zhouhang2991@gmail.com (H. Zhou); zhangwm@ustc.edu.cn (W. Zhang); ynh@ustc.edu.cn (N. Yu)
0000-0001-5553-9482 (J. Lin); 0000-0002-4703-4641 (W. Zhou); 0000-0001-9294-9624 (H. Liu); 0000-0001-7860-8452 (H. Zhou); 0000-0001-5576-6108 (W. Zhang); 0000-0003-4417-9316 (N. Yu)

against physical changes in different videos, *e.g.*, volume and face angle. This precisely provides a distinguishing feature for forgery detection. By selecting the phonemes with the top-5 deviations, we integrate the corresponding 5 well-trained sub-models into an ensemble model for maximizing the discriminability of real and fake videos.

To verify the effectiveness, we have conducted extensive experiments on both the public DFDC dataset and a self-organized lip forgery video dataset which contains four sub-datasets. The experimental results demonstrate that our method outperforms the current state-of-the-art detection methods on cross-dataset evaluation and multiple class classification. In addition, our method is also competitive on single dataset classification.

- We propose a multi-phonemes selection based framework for lip forgery detection task, which takes full advantage of the visual and aural information in lip forgery videos.
- We establish 12 categories of phoneme-lip mapping relationships and explore the robustness between the open-close amplitudes on each pair for real/fake classification. We also organize a new lip forgery dataset which is helpful to facilitate the development of lip forgery detection methods.
- Extensive experiments demonstrate that our method outperforms state-of-the-art approaches for lip forgery detection on both the public DFDC dataset and a self-organized lip forgery dataset.

2. Related work

2.1. Deep Face Forgery

According to different forgery regions, existing methods can be divided into two categories: *full-face synthesis* and *local region forgery*. Full-face synthesis usually synthesizes a whole source face and swaps it to the target. Typical works are [4, 14].

Local region forgery is a more common type, focusing on slight manipulation of partial facial regions, *eg*, eyebrow locations and lip shapes. *Lip-sync* [5] is able to modify the lip shapes in Obama's talking videos to accurately synchronize with a given audio sequence. [15] leverages 3D modeling for specific face videos to make the control of lip shapes more flexible. *First Order Motion* [16] uses video to drive a single source portrait image to generate a talking video. The detection of local region forgery is more challenging due to the subtle and local nature.

2.2. Face Forgery Detection

Early works explored visual artifacts, *eg*, the abnormality of eye blinking and teeth. Learning-based detection methods have become mainstream in very recent years. [7] uses XceptionNet [17] to extract features from spatial domain. F^3-Net [9] achieves state-of-the-art using frequency-aware decomposition. However, since the audios are lacking in most public deepfake datasets, these methods are designed in a universal manner with no consideration of audios matching. They perform well in full-face synthesis detection but is not adequate to recognize the subtle artifacts in local region forgery.

Recently, [11, 12] utilize Siamese network to calculate the feature distances in multi-modalities. If manipulation is conducted on a small segment of the video, this will weaken the inconsistency among these modalities at the video level, leading to a decrease in detection performance.[10] establishes one single phoneme-viseme mapping for a specific person, which severely restricts the application scenario. To address the above limitations, we propose a multi-phoneme selection based framework for lip forgery video detection.

3. Method

In this section we will elaborate the multi-phoneme selection based framework. Before that, an important observation of lip forgery will be introduced first.

3.1. Motivation

Lip forgery modifies a specific person's lip shape to match arbitrary audio contents, thus establishing a close relationship between them. However, due to imperfections in the manipulation, uncontrollable artifacts may be generated to hinder the matching.

As shown in Figure 1, when saying the word "apple", the lips in the forgery videos are more blurred to open well. Although this nuance is not easy to perceive by human eyes, a well-designed detector can capture it. Nevertheless, the lip shape itself fluctuates in a certain range under different expressions, large fluctuation indicates poor robustness.

Based on this observation, it is necessary to establish a robust mapping from audios to lip shapes. Inspired by recent works in Audio-Visual Speech Recognition [18], we divide all audio contents into 12 phonemes categories as the smallest identifiable units. Each phoneme set consists of various vowels, consonants and quiet soundmark, which can be used to train sub-model independently to distinguish real/fake lips. Eventually, we select several sub-models to integrate the final classifier considering the trade-off between efficiency and performance. The framework is depicted in Figure 2.

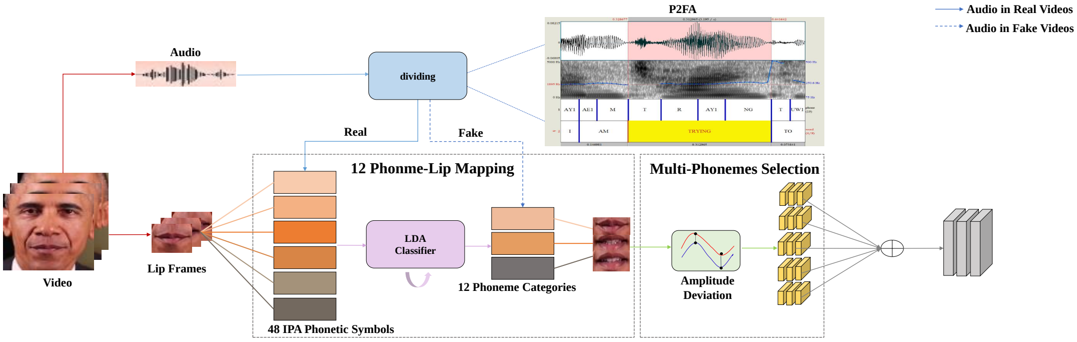


Figure 2: The framework of ours. Through 12 phoneme-lip shape mapping and multi-phonemes selection, we obtain the final ensemble detection model.

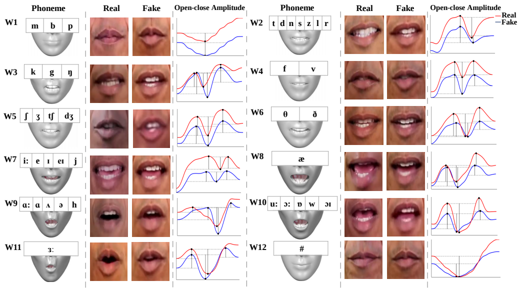


Figure 3: Illustration of the robust phoneme categories. We exhibit the basic lip patterns with similar phonetics, visually compare the real and fake lip shapes and the average open-close amplitude curves.

3.2. Correlations Establishment from Phonemes to Lip shapes

For a given talking video, we use OpenFace [19] to align each frame and crop the lip area to 128×128. These lip images will be categorized into different phoneme set and used as training/testing data for real/fake classification.

To establish the mapping from phonemes to lip shapes, we first process all the real videos. According to the International Phonetic Alphabet (**IPA**) we divide the lip shapes into 48 classes. For a given lip shape, we calculate the Mahalanobis distance d_c of the open-close amplitude between the current lip shape $\mathbf{x}$ and mean $\mathbf{x_c}$ of each class.

$$d_c(\mathbf{x}) = \sqrt{(\mathbf{x} - \bar{\mathbf{x}}_c)^T \cdot \Sigma_c^{-1} \cdot (\mathbf{x} - \bar{\mathbf{x}}_c)} \qquad (1)$$

Next, we estimate the probabilities of it belonging to each class, and assign the sample to the class with the highest normalized probability P_c:

$$P_c(\mathbf{x}) = \frac{p(c \mid \mathbf{x})}{\sum_{c=1}^{C} p(c \mid \mathbf{x})} \qquad (2)$$

Here, $p(c \mid \mathbf{x})$ is the probability of $\mathbf{x}$ belongs to class c, which is computed as the ratio between the in-class and the out-of-class distribution from the previous distance d_c, following the Gaussian distribution with means $\mu_c, \mu_{\bar{c}}$ and variances $\sigma_c, \sigma_{\bar{c}}$, respectively :

$$p(c \mid \mathbf{x}) = \frac{1 - \Phi\left(\frac{d_c(\mathbf{x}) - \mu_c}{\sigma_c}\right)}{\Phi\left(\frac{d_c(\mathbf{x}) - \mu_{\bar{c}}}{\sigma_{\bar{c}}}\right)} \qquad (3)$$

After obtaining the mapping, a multi-class LDA classifier pre-trained on [20] is utilized for classification. However, different classes may share the same lip shape appearance, *e.g.*, *m,b,p*. By iteratively merging similar phonetic symbol classes, we obtain 12 distinguishable real lip shapes named "phoneme" (from W1 to W12) with robustness. A visual example is given in Figure 3.

In fake videos, the lip shapes have been manipulated. As illustrated in Figure 1, the opening amplitudes of fake lips are quite different from real ones, thus directly using the phoneme classifier trained on real lips may lead to misclassification. Since the audio contents in fake videos are not modified, we decide to use them as the guidance for fake lips classification. First, Google's Speech-to-Text API is used to obtain the corresponding transcribed texts from the audios. Both the texts and audios are then fed into the P2FA toolkit [21]. By conducting forced alignment on phonemes and words, we get the start and end time for each phoneme, the lip images during this period will be categorized into the current phoneme. In Figure 2, the P2FA section clearly shows the alignment procedure.

3.3. Multiple Phonemes Selection

Although the lip shapes in one phoneme set are similar, the open-close amplitudes among phonemes are quite different. We use dlib 68 face landmarks detector [22] to compute the vertical axis value between the 63th and 67th landmarks: D = $(y_{63}$-$y_{67})$. Here D represents the

Table 1
Amplitude Deviation Values for 12 phonemes in self-organized dataset. The Top-5 phonemes with the largest amplitude deviation for each sub-dataset are in bold.

Forgery Methods	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12
Obama Lip-sync[5]	**33.00**	**31.13**	21.63	**33.12**	**34.87**	27.625	**37.50**	24.37	26.87	24.00	22.38	25.25
Audio Driven[15]	15.00	**23.62**	18.50	**26.62**	**28.00**	**25.50**	**29.50**	20.63	17.37	18.25	17.00	12.50
First Order Motion[16]	25.13	23.75	**34.67**	**37.12**	**34.87**	22.50	23.38	25.125	**33.50**	**29.50**	21.75	20.88
Wav2lip[6]	**35.51**	**34.71**	26.71	28.01	25.12	25.43	**35.12**	28.76	27.32	**33.84**	29.96	**33.60**

open-close amplitude of the current lip shape. Using the number of frames as the horizontal axis, we calculate D for each frame during the period of the phoneme. In Figure 3, we plot two average amplitude curves for each set, the red curves represent the real videos while the blue for fake.

In W1 and W2, the real and fake curves are widely separated with almost no overlap, while in W3 and W6, there are partially stacked areas. This observation indicates that the real and fake lips are more discriminative in certain phoneme sets. To select the most distinguishable phonemes W for classification, we calculate the differences between the maximum and minimum values $D_{W_{max}}$,$D_{W_{min}}$ of real/fake curves, respectively. We define the amplitude deviation D_W to represent the discrepancy between real and fake in each phoneme W: $D_W = \frac{1}{2}(D_{W_{max}} + D_{W_{min}})$.

Considering the potential differences in forgery methods, the amplitude deviations of a single phoneme are not identical. As listed in Table 1, the phonemes with top-5 amplitude deviations are in bold, and we will introduce the self-organized dataset in Section 4.

3.4. Sub-classification Models training and Ensemble

After selecting the phoneme-lip sets for each forgery method, we train sub-classification models based on them. Each sub-model can be used independently for real/fake lips discrimination. Here we adopt XceptionNet [17] as the backbone and transfer it to our task by resizing the input to 128×128 and replacing the final connected layer with two outputs.

To obtain a stronger detection performance, we integrate the sub-models into an ensemble one. The average weight for each is equal to ensure the contribution is maximized. Furthermore, phoneme units in the video will last for some duration, which contain several lip frames. Both the lip frame numbers f and sub-models N will influence the detection accuracy of the final ensemble model, hence we experiment on them respectively. The results in Section 4 demonstrate that when $f = 4$ and $N = 5$, the ensemble model can achieve excellent performance without importing extra complexity.

Table 2
The composition of our self-organized dataset, including the numbers of videos and frames. The whole dataset consists of four sub-datasets.

Dataset	Real/Fake	Total	Frames
Obama Lip-sync[5]	28	56	62534
Audio Driven[15]	24	48	54416
First Order Motion[16]	24	48	53614
Wav2lip[6]	28	56	63736

4. Experiments

In this section, we initially introduce a new lip forgery video dataset organized by this paper. Several parameter studies can verify the optimality of our settings. Further experiments are provided to demonstrate the effectiveness of our proposed framework on DFDC and self-organized dataset, as well as the transferability between them.

4.1. Public Dataset and New Lip Forgery Dataset

Many datasets [7, 23] have been public for deepfake detection task. Although with large scale and various forgery methods, most fake videos do not contain the audios, which still tampered in a full-face synthesized manner. So far, there is no dedicated dataset released for lip forgery detection. In this paper, we use one public audio-visual deepfake dataset and organize a new dataset targeting the lip forgery detection task.

Public DFDC Dataset [24] has been published in the Deepfake Detection Challenge, using multiple manipulation techniques and adding audios to make the video scenarios more natural. To make a fair comparison, we align with the settings of [11], using 18,000 videos in the experiments.

New Lip Forgery Dataset To build the new lip fogery dataset, we adopt four state-of-the-art methods [5, 15,

Table 3
Parameter study of frame selection. $f = 4$ can guarantee the best performance and avoid the overlap with other phonemes.

Frame Numbers	$f = 3$	$f = 4$	$f = 5$	$f = 6$	$f = 7$	$f = 8$
ACC (%)	96.21	**97.73**	96.21	96.97	97.73	97.73
AUC (%)	97.45	**98.89**	97.45	97.83	98.89	98.89

16, 6] to generate fake videos. The composition of the organized dataset is elaborated in Table 2.

4.2. Experimental Settings

As mentioned before, XceptionNet is the baseline. According to the particularity of the public DFDC dataset and self-organized dataset, we adopt different training strategies. On the large DFDC dataset, we train our model with a batch size of 128 for 500 epochs. Due to the distinctly smaller size of the self-organized dataset, we train with a batch size of 16 for 100 epochs on each sub-dataset. For both datasets, we uniformly use the Adam optimizer with the learning rate of 0.001 and employ ACC (accuracy) and AUC (area under ROC curve) as evaluation metrics.

4.3. Parameter Study

Frame Selection. As showed in Figure 2, a single phoneme unit will include several lip frames. We use f to represent the number of lip frames, the value of f has an impact on the competence of the model. Few lip frames result in missing lip features of the current phoneme, while extra frames may overlap with others.

In order not to introduce disturbances from other factors, we experiment on the Obama Lip-sync dataset. We integrate all the 12 phoneme sub-models into one and take the beginning time of each phoneme as the center to select the surrounding frames f. Table 3 displays the accuracy of f from 3 to 8. The accuracy reaches 97.73% when $f = 4$, 7 and 8. Considering the tradeoff between accuracy and complexity, we finally choose $f = 4$.

Phoneme Selection. Still executing on the Obama Lip-sync dataset, we use N to denote the number of selected phonemes. Referring to the amplitude deviations ranking listed in Table 1, we integrate the sub-models from 2 to 12, the highest accuracy is achieved when $N = 5$. Thus we choose phoneme sets with the top 5 amplitude deviations to train sub-models.

4.4. Evaluation on DFDC Dataset

In this section, we compare our method with previous deepfake detection methods on DFDC. The ratio of training and testing sets is 85:15. Even though we only crop the lip region of the face, we still achieve a competitive performance. In Table 4, our method achieves 91.6% on AUC, which outperforms not only the vision based full-face method but also the audio-visual based multi-modal method. Among them, Syncnet[12] detects the synchronization from audios to video frames, achieves 89.50% on AUC, while ignoring the content matching between them. The improvement in ours mainly benefits from the establishment of the phoneme-lip mapping, where the selected phonemes W2,W5,W7,W10 and W11 are robust to various external disturbances in DFDC such as face angle, illumination, and video compression, boosting the detection capability of the ensemble model.

Table 4
Comparison of our method(Xception) with other techniques on the DFDC dataset using the AUC metric. We select sub-models of W2, W5, W7, W10, and W11 for integration, and our result is competitive against Syncnet and Siamese-based methods.

Methods	DFDC	Modality
Xception-c23[17]	72.20	Video
Meso4[25]	75.30	Video
DSP-FWA[26]	75.50	Video
MBP[10]	80.34	Audio & Video
Siamese-based[11]	84.40	Audio & Video
Syncnet[12]	89.50	Audio & Video
Ours (Xception)	**91.60**	Audio & Video

Moreover, we respectively visualize the Gradient-weighted Class Activation Mapping (Grad-CAM) [28] for the baseline and ours, as shown in Figure 4. It shows that our method can significantly include the surrounding regions such as the upper and lower lips, which facilitates the network to focus on the open-close amplitudes and is in line with our motivation. In contrast, the baseline model mainly concerns the internal teeth regions, losing the edge information.

4.5. Evaluation on Self-organized Dataset

In this section, we conduct experiments on self-organized dataset to verify the performance of real/fake classification and multiple classification.

4.5.1. Evaluation of Real/Fake Classification

For each sub-dataset, We use different phonemes to integrate the final classification model, the selections are listed in Table 5. The baseline model (Xception) is directly trained on all continuous frames of real/fake videos. Further, to verify that our method is not restricted by the backbone, we adopt another network architecture ResNet-50 [29] which performs well in image classifica-

Table 5
Evaluation of Real/Fake Classification. For each dataset, the performance of our approach surpasses baselines (Xception/ResNet-50) and existing state-of-the-art detection methods.

Methods	Obama Lip-sync[5] (W1-W2-W4-W5-W7)		Audio Driven[15] (W2-W4-W5-W6-W7)		First Order[16] (W3-W4-W5-W9-W10)		Wav2lip[6] (W1-W2-W7-W10-W12)	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
MBP[10]	93.54	96.03	-	-	-	-	-	-
Siamese-based[11]	90.53	93.01	87.47	89.86	92.03	95.21	84.77	88.64
Syncnet[12]	92.18	95.21	90.83	92.89	92.18	95.56	86.08	90.16
ResNet-50	79.38	85.72	68.65	72.62	86.97	89.40	75.23	78.96
Xception[17]	84.82	89.19	70.18	78.43	88.83	93.71	78.54	80.78
Ours(ResNet-50)	96.35	97.67	94.67	96.40	96.25	97.62	95.12	96.74
Ours(Xception)	**97.73**	**98.89**	**95.84**	**97.61**	**97.59**	**98.60**	**96.43**	**97.89**

Table 6
Evaluation of multiple classification. In the table, except for the average AUC (%) in the last column, other data represent the ACC (%). Here, Our method integrates the sub-models of W2, W3, W4, W7 and W8 into the ensemble one, which largely outperforms the advanced methods.

Methods	Real	Obama Lip-sync[5]	Audio Driven[15]	First Order[16]	Wav2lip[6]	Average ACC	Average AUC
Siamese-based[11]	92.91	77.63	70.86	85.14	79.44	81.20	88.45
Syncnet[12]	94.89	78.79	74.33	88.62	81.54	83.46	90.53
Xception[17]	92.13	73.44	55.13	78.01	77.27	75.37	83.12
Ours (Xception)	**96.21**	**95.96**	**87.50**	**96.97**	**94.88**	**94.29**	**96.84**

tion tasks. The results in Table 5 demonstrate that our method outperforms the previous methods, where MBP is designed for Obama lip forgery and the Audio Driven dataset is challenging with low video resolution and the blocking of microphones or arms.

4.5.2. Evaluation of Multiple Classification

To further distinguish different forgery methods, in the 4 sub-datasets, we label all real lips with 0 and fake lips with 1 ∼ 4 individually. W2, W3, W4, W7, W8 are chosen to train the classification model.

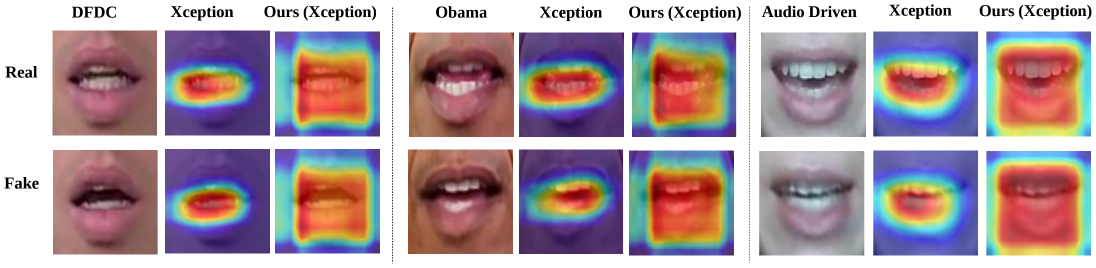
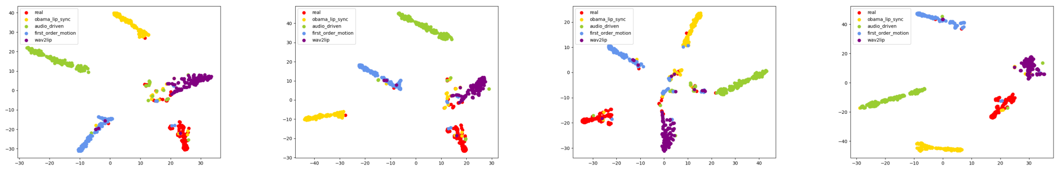


Figure 4: The Grad-CAM of the baseline Xception and ours, including DFDC dataset and two forgery methods in self-organized dataset. Ours can easily capture more lip regions.



(a) Siamese-based (b) Syncnet (c) Xception (d) Ours(Xception)

Figure 5: Feature distributions visualization from Siamese-based (a) to ours (d) on multiple classification. In the four methods, ours contains less outliers and widely separates the real and fake classes.

Table 7
Evaluation on cross-dataset. The testset is self-organized dataset. Ours (W2,W5,W7,W10,W11) achieves better results.

Methods	ACC	AUC
MBP[10]	57.94	59.12
Siamese-based[11]	59.51	60.68
Syncnet[12]	60.11	61.79
ResNet-50[27]	54.74	57.67
Xception[17]	56.80	58.89
Ours (ResNet-50)	62.38	63.51
Ours (Xception)	**63.67**	**64.05**

Table 6 verifies that the ensemble model can be applied to multiple classification scenarios. We also intuitively visualize the t-SNE[30] feature distributions from Siamese-based to ours. As shown in Figure 5, our method is superior to find latent dissimilarity in high-dimensional space with fewer outliers.

4.6. Evaluation on cross-dataset

Transferability is evaluated by training on DFDC but testing on self-organized dataset where all lips are labeled as real/fake. Table 7 shows better transferability of ours in detecting universal artifacts in various datasets.

5. Conclusion

Lip forgery detection is an extremely challenging task in deepfake detection due to the subtle and local modifications. In this paper, we present a multi-phoneme selection based framework. Varying from existing deepfake detection, it takes full advantage of the particularity of lip forgery videos, establishing a robust mapping from audio to lip shapes. 12 categories of phonemes are determined as the smallest identifiable unit for various lip shapes and the phonemes with top-5 distinguishability are selected to train sub-classification models. In addition, we organize a new dataset consists of four sub-datasets, which is the first one organized for lip forgery detection task. Extensive experiments demonstrate the effectiveness of our framework, including the challenging task of cross-dataset evaluation.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047, U1636201, 62002334, by the Anhui Science Foundation of China under Grant 2008085QF296, by the Exploration Fund Project of the University of Science and Technology of China under Grant YD3480002001 and the Fundamental Research Funds for the Central Universities under Grant WK2100000011.

References

[1] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2387–2395.

[2] Y. Nirkin, Y. Keller, T. Hassner, Fsgan: Subject agnostic face swapping and reenactment, 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 7183–7192.

[3] DeepFakes, Deepfakes github, http://github.com/deepfakes/faceswap, 2017. Accessed 2020-08-18.

[4] FaceSwap, Faceswap github, http://https://github.com/MarekKowalski/FaceSwap, 2016. Accessed 2020-08-18.

[5] I. K.-S. Supasorn Suwajanakorn, Steven Seitz, Synthesizing obama: Learning lip sync from audio, SIGGRAPH 36 (2017) 95.

[6] R. PrajwalK, R. Mukhopadhyay, V. Namboodiri, C. Jawahar, A lip sync expert is all you need for speech to lip generation in the wild, Proceedings of the 28th ACM International Conference on Multimedia (2020).

[7] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, arXiv preprint arXiv:1901.08971 (2019).

[8] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face x-ray for more general face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5001–5010.

[9] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, in: ECCV, 2020.

[10] S. Agarwal, H. Farid, O. Fried, M. Agrawala, Detecting deep-fake videos from phoneme-viseme mismatches, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 2814–2822.

[11] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emotions don't lie: A deepfake detection method using audio-visual affective cues, ArXiv abs/2003.06711 (2020).

[12] K. Chugh, P. Gupta, A. Dhall, R. Subramanian, Not made for each other- audio-visual dissonance-based deepfake detection and localization, Proceedings of the 28th ACM International Conference on Multimedia (2020).

[13] H. L. Bear, R. Harvey, Phoneme-to-viseme mappings: the good, the bad, and the ugly, ArXiv abs/1805.02934 (2017).

[14] L. Li, J. Bao, H. Yang, D. Chen, F. Wen, Faceshifter: Towards high fidelity and occlusion aware face swapping, arXiv preprint arXiv:1912.13457 (2019).

[15] R. Yi, Z. Ye, J. Zhang, H. Bao, Y. Liu, Audio-driven talking face video generation with natural head pose, ArXiv abs/2002.10137 (2020).

[16] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe, First order motion model for image animation, ArXiv abs/2003.00196 (2019).

[17] F. Chollet, Xception: Deep learning with depthwise separable convolutions, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 1800–1807.

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

[19] T. Baltrusaitis, P. Robinson, L.-P. Morency, Openface: An open source facial behavior analysis toolkit, 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (2016) 1–10.

[20] A. Ortega, F. Sukno, E. Lleida, A. Frangi, A. Miguel, L. Buera, E. Zacur, Av@car: A spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition, in: LREC, 2004.

[21] S. Rubin, F. Berthouzoz, G. J. Mysore, W. Li, M. Agrawala, Content-based tools for editing audio stories, in: UIST '13, 2013.

[22] D. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.

[23] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216.

[24] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge dataset, arXiv preprint arXiv:2006.07397 (2020).

[25] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (2018) 1–7.

[26] Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[30] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008) 2579–2605.