CONTENT-INSENSITIVE DYNAMIC LIP FEATURE EXTRACTION FOR VISUAL SPEAKER AUTHENTICATION AGAINST DEEPFAKE ATTACKS

Zihao Guo, Shilin Wang, Senior Member, IEEE*

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China

ABSTRACT

Recent research has shown that lip-based speaker authentication system can achieve good authentication performance. However, with emerging deepfake technology, attackers can make high fidelity talking videos of a user, thus posing a great threat to these systems. Confronted with this threat, we propose a new deep neural network for lip-based visual speaker authentication against human imposters and deepfake attacks. One dynamic enhanced block with context modeling scheme is designed to capture a user's unique talking habit by learning from his/her lip movement. Meanwhile, a cross-modality content-guided loss is designed to help extract discriminative features when learning from different lip movement of a user uttering different content. This loss makes the proposed method insensitive to content variation. Experiments on the GRID dataset show that the proposed method not only outperforms three state-of-the-art methods but also simplifies the training process and reduces the training cost.

Index Terms— Dynamic feature extraction, visual speaker authentication, deepfake attacks, contrastive learning

1. INTRODUCTION

Lip features have been used for speaker authentication (known as Visual Speaker Authentication, VSA) [1, 2, 3, 4, 5, 6, 7] since 1990s. However, With the development of deepfake, high quality talking face videos of a speaker can be easily forged with his/her visual corpus using either face swapping or face reenactment [8, 9, 10]. In some deepfake methods [11, 12, 13], even model training is unnecessary and pre-trained model can be directly used to generate deepfake videos of a speaker. The deepfake has already threatened traditional face-based authentication systems [14] and lip-based authentication systems are under threat as well. To attack these systems, imposters/attackers can record videos of themselves uttering the random password, and generate deepfake videos of the target user uttering the same random password with the user's visual corpus.

Confronted with the new threat of deepfake, many deepfake detection methods based on manipulation artifacts have been proposed [8, 9, 10], but many of them lack the generalization ability and can only be used to detect certain types of deepfake forgeries [15, 16, 17, 18]. However, various types of deepfake videos of a user can be made to attack authentication systems. In the recent work [18, 19], two deep neural networks based on dynamic lip features were proposed. They can achieve good authentication results and defend against both human imposters and various deepfake attacks in random prompt text scenario for VSA. However, they suffer from decline in model's performance when learning from different lip movement of a user uttering different content. As a result, for random prompt text with many different words spoken, they have to train many word-level models, resulting in troublesome training process and large training cost.

Faced with problems mentioned above, we propose a new lip-based VSA method. The main contributions of our work are three-folds: First, a new deep neural network aimed at solving the content-sensitive problem and improving the sentence-level authentication performance is proposed. Second, a novel block with context modeling mechanism is designed to extract more discriminative identity-related dynamic features from a user's lip movement. Third, a content-guided loss combining both visual and text modality is designed to capture talking habit under different content. By adopting this loss, for each speaker, only one model is needed to be trained instead of many word-level models in random prompt text scenario for VSA. Hence, the training cost is reduced and the training process is simplified.

2. CHALLENGE AND MOTIVATION

Established methods [18, 19] suffer from the content-sensitive problem. They can extract discriminative dynamic features only when dealing with talking videos of a user uttering the same speech content, such as a fixed word. As a result, many word-level models have to be trained to do sentence-level VSA.

It is assumed that this problem results from the complexity of dynamic feature space caused by extra content information. It is hard for the model to extract discriminative identity-related dynamic features (talking habit) under differ-

*Shi-Lin Wang is the corresponding author. This work was supported by the National Natural Science Foundation of China (62271307, 61771310) and Key RD Program Major (Key) Project of Science and Technology Department of Ningxia.

ent content compared with fixed content. As a result, the extracted features are not discriminative enough to defend against deepfake attacks. To solve the problem, we resort to contrastive learning to correlate word-level video embeddings with their related word embeddings. This can be seen as adding extra constraint to the model by dividing the dynamic feature space into many word-level subspaces based on the content of different words and then extracting discriminative identity-related features in each subspace to bring down the complexity caused by the content information. This loss should help capture talking habit under different content of words, thus making the proposed method robust to content variation. For talking video of the user uttering every word, the model should be able to extract discriminative identity-related features. As a result, only one model is needed for a user instead of many word-level models, which simplifies the training process and reduces the training cost.

3. PROPOSED METHOD

In this section, the overall architecture of the proposed Content Insensitive Dynamic Enhanced Network (CIDE-Net) is introduced. Meanwhile, the Dynamic Enhanced block (DE-block) and the content-guided loss are elaborated.

3.1. Overall Architecture

The overall architecture of the proposed neural network is shown in Fig.1. In the preprocessing stage, following [19], we use the Penn Phonetics Lab Forced Aligner [20] to get the timestamp of each word (word alignment) with the input audio and the provided random prompt text. Meanwhile, the Dlib detector [21] is used to cut the mouth region of the input video. The word alignment is used to separate the input video into word-level video segments during the training stage. Then two stacked dynamic enhanced blocks are used to process these word-level video segments to extract dynamic features from lip movement. Features maps extracted from the first layer and second layer are concatenated together as the final dynamic features and this helps combine features from both shallow layer and deep layer. The dynamic features will then be processed by a 3D Convolutional Neural Network (C3D) [22] to get the representations of all the single word-level video segments ($N \times D$, N denotes the number of words in a sentence and D denotes the dimension of representation of a word-level video segment).

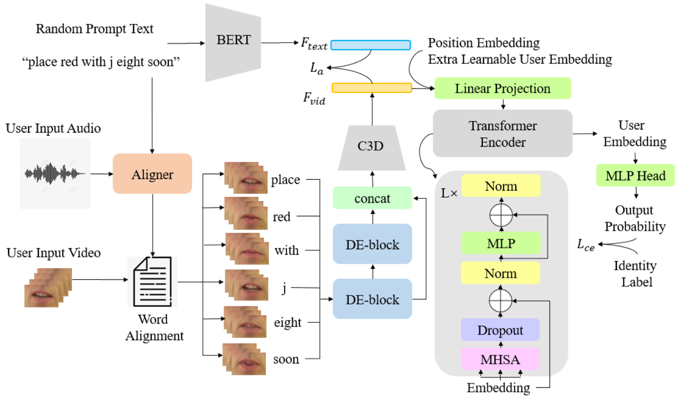

Fig. 1. The overall architecture of the proposed CIDE-Net.

Similar to [23], we introduce a linear projection layer and a transformer encoder to fuse all the talking habits reflected in word-level video segments of the user to a sentence-level user embedding. MHSA denotes multi-head attention layer [24] and MLP denotes Multilayer Perceptron in Fig.1. At last, the user embedding is fetched and passed to a MLP head to get the output probability, which measures the credibility of an authentication request that whether the request is sent from who he/she claims to be.

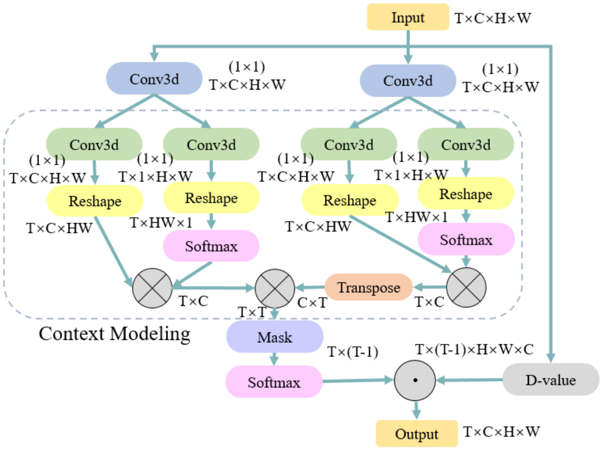

Fig. 2. Architecture of dynamic enhanced block, where ⊙ denotes weighted sum of corresponding terms.

3.2. Dynamic Enhanced Block

This block is inspired by [19], in which differences between frames are calculated and correlations between different frames are considered. However, compared with simple Global Average Pooling (GAP) operation in the calculation of correlations between frames in [19], we take a fine-grained modeling scheme inspired by context modeling in [25]. By computing a global attention map and sharing this global attention map for all query positions, we can model global context as a weighted average of the features at all positions. This helps us get fine-grained channel features of one frame compared with GAP, thus obtaining more discriminative dynamic features by use these fine-grained features for correlation calculation between frames.

The detail of this block is shown in Fig.2. The definition of the block with the operation of mask and D-value is similar to the diff-block in [19].

3.3. Content-Guided Loss

Dynamic lip features extracted from different word-level video segments contain both content and identity information (talking habit) and the extra content information add the complexity of dynamic feature space. As a result, it is hard for the model to extract discriminative identity-related dynamic features under different content compared with fixed content.

Confronted with this problem, we find that When different people speak the same word, their talking habits are different but speech content is the same. So it can be expected that the video feature embeddings of the same word spoken by different people are close to the related word embedding compared with other words. As for a user speaking different words, each video feature embedding of the user uttering these words is expected to be close to related word embedding since they share the same content.

Based on the above consideration [26], we resort to contrastive learning and a content-guided loss which is in the form of InfoNCE loss[27] is designed to correlate the word-level video embedding with related word embedding by making the former close to the latter. Meanwhile, the word-level video feature embeddings are pushed away from their unrelated word embedding.

This helps to add extra constraint to the model by dividing the dynamic feature space into many word-level subspaces and extracting discriminative identity-related features in each subspace to bring down the complexity caused by the content information. This loss helps capture talking habit under different content and enable the model to extract discriminative word-level dynamic features regardless of content variation. For videos of the user uttering every single word, the model which is robust to content variation, can extract dynamic features reflecting the user's talking habit to defend against deepfake attacks. A discriminative sentence-level video feature can finally be obtained by fusing all these word-level video features together. As a result, only one model is needed for a user instead of many word-level models, which simplifies the training process and reduces the training cost.

For feature distances measurement, we adopt the cosine distance where closer features render larger scores. The content-guided loss can be formulated and shown in (1), where N denotes number of words in a sentence as mentioned above and F_V^i and F_w^i denotes video feature and text feature of the i-th word respectively. The random prompt text is processed by a pre-trained BERT [28] to get the word embeddings during the training stage.

$$L_a = \sum_{i=1}^{N} -log[\frac{e^{D(F_v^i, F_w^i))}}{e^{D(F_v^i, F_w^i)} + \sum_{j=1}^{N-} e^{D(F_v^i, F_w^j)}}] \qquad (1)$$

3.4. Implementation Details

A cross-entropy loss (L_{ce}) is calculated between output probability and user label to discriminate a specific user with other users. Meanwhile, a content-guided loss (L_a) is calculated between word embeddings and word-level video segment embeddings. In the training stage, random prompt text, word alignment and user's video are used as input and the whole CIDE-Net is optimized by the addition of content-guided loss and cross entropy loss.

It is noticeable that the weights of BERT [28] are frozen during training and any prior knowledge about deepfake is not required for training. Following [18, 19], for a specific speaker in the user set, this user's videos are seen as positive samples and videos of all the other users are seen as negative samples in the training stage. The threshold θ for test is obtained when the Equal Error Rate (EER) is reached in the evaluation stage. In the test stage, when the output probability is greater/smaller than the threshold θ, the video is recognized as a user/imposter sample. Each user need a trained model and test results are averaged over all the users.

4. EXPERIMENTS

4.1. Experiment Setup

We use the GRID dataset[29] to validate the effectiveness of the proposed CIDE-Net following [18, 19]. The experiment setting follows [18] in sentence level. Twenty-four speakers are randomly selected as the user set and the remaining eight speakers are selected as the attacker set. Each speaker in the attacker set and each speaker in the user set form a pair, resulting in 24×8=192 pairs in total. For each user-attacker pair, four kinds of deepfake videos of the user are made by either the attacker's audios or videos to attack, including faceswap[1], faceswap-gan[2], wav2lip[11], wav2lip-gan[11] (fs,fg,ls,lg for short).

Area Under Curve (AUC), False Rejection Rate (FRR) and Half Total Error Rate (HTER) are used as evaluation metrics. FRR denotes the false rejection rate. FAR denotes the false acceptance rate. HTER denotes the half total error rate. HTER is the average of FRR and FAR.

4.2. Comparisons With State-of-the-Art VSA Methods

To investigate the effectiveness of the proposed CIDE-Net, three state-of-the-arts [3, 18, 19] are adopted for comparison and the proposed method outperforms them in HTER$_{fs}$, HTER$_{fg}$ and HTER$_{lg}$, as shown in Table 1. This is mainly because the dynamic enhanced block can better capture the discriminative dynamic features of the speaker in word-level and a transformer can fuse all these word-level features into more discriminative sentence-level dynamic features.

Note that only one model is needed to be trained for a user in the proposed method. For Yang's and Ma's, however

[1] https://github.com/deepfakes/faceswap

[2] https://github.com/shaoanlu/faceswap-GAN

22 word-level authentication models have to be trained respectively with word video segments. In the test stage, for a random prompt text sentence with many words, these models are used to get word-level authentication results and a final sentence-level authentication result is obtained by voting [18]. Compared with their methods, training process is simplified and training cost is reduced as well in the proposed method.

Table 1. Authentication results comparisons (in %) with the state-of-the-arts.

Model	FRR	HTER$_{hm}$	HTER$_{fs}$	HTER$_{fg}$	HTER$_{ls}$	HTER$_{lg}$
LOCP	**0.20**	**0.40**	12.80	22.50	14.40	17.35
Yang's	0.40	0.60	3.30	3.50	2.90	5.96
Ma's	1.57	2.10	8.43	2.29	**2.08**	4.70
Proposed	1.65	0.84	**1.33**	**0.84**	2.62	**4.66**

4.3. Content-Insensitivity of the Proposed Method

To show the content sensitive problem of previous methods and the content-insensitivity of the proposed method, we train only one model instead of 22 word-level models for Yang's and Ma's to extract dynamic features as comparisons. As can be seen in the Table 2, the defense ability of two state-of-the-arts [18, 19] drops a lot compared with results in the Table 1 when learning from different lip movement of a user uttering different content.

Meanwhile, due to the content-guided loss introduced in the training stage, the proposed model is able to extract more discriminative identity-related features to defend against deepfake attacks under different content in the test stage, compared with their methods.

Table 2. Authentication results comparisons (in %) with the state-of-the-arts.

Model	FRR	HTER$_{hm}$	HTER$_{fs}$	HTER$_{fg}$	HTER$_{ls}$	HTER$_{lg}$
Yang's	0.46	**0.24**	22.60	8.32	33.25	28.21
Ma's	**0.38**	1.16	15.65	6.40	8.60	21.90
Proposed	1.65	0.84	**1.33**	**0.84**	**2.62**	**4.66**

4.4. Comparisons With State-of-the-Art DeepFake Detection Methods

Table 3 shows the detection results on unseen attacks between the proposed method and state-of-the-art deepfake detection methods [15, 16, 30, 31], especially those newly proposed methods which have strong generalization ability. Their pre-trained models are used for test and the detection results show that the proposed method outperforms their methods on unseen attacks. The superiority of the proposed method is mainly because it only extracts biometric features from real samples and does not rely on deepfake samples, thus having strong defense ability on unseen deepfake attacks.

4.5. Ablation Study

The results of ablation studies are shown in the Table 4. When learning from videos of a user uttering prompt text with words of different content, content-guided loss (CG-loss) bring down the complexity of the dynamic feature space by dividing it into many word-level subspaces. This helps the model better extract discriminative word-level identity-related features under different content and finally obtain discriminative sentence-level features. Compared with the proposed model, the defense ability of the model without CG-loss drops a lot.

Table 3. Comparison results (in %) with the state-of-the-arts of deepfake detection methods.

Model	AUC$_{fs}$	AUC$_{fg}$	AUC$_{ls}$	AUC$_{lg}$
FRDM	77.73	76.19	82.58	80.94
MAT	82.76	88.31	91.19	92.73
LipForensics	73.39	81.31	87.28	90.36
SBI	70.44	66.08	68.50	70.44
Proposed	**99.85**	**99.91**	**99.71**	**99.45**

Table 4. Results (in %) of ablation studies.

Model	FRR	HTER$_{hm}$	HTER$_{fs}$	HTER$_{fg}$	HTER$_{ls}$	HTER$_{lg}$
w/o CG-loss	**1.50**	2.15	5.03	3.30	15.66	21.56
w/o DE-block	1.56	3.92	5.42	2.62	25.71	28.01
w/o CM	2.19	1.11	3.08	1.33	7.17	12.98
CIDE-Net	1.65	**0.84**	**1.33**	**0.84**	**2.62**	**4.66**

Meanwhile, dynamic enhanced block (DE-block) can capture a user's unique talking habit to defend again human imposters and deepfake attacks. Compared with normal GAP, the Context Modeling scheme (CM) of the DE-block also plays an important role by extracting more discriminative identity-related features, thus obtaining lower FRR and FAR at the same time when dealing with prompt text with words of different content.

5. CONCLUSIONS

The main contribution of our work is to solve the content-sensitive problem existed in the SOTA VSA approaches and to improve the sentence-level authentication performance. By adopting the newly proposed content-guided loss, only one model is needed to be trained for each speaker instead of many models for all the words in the vocabulary in random prompt text scenario for VSA. Hence, the training cost is reduced and the training process is simplified. On the other hand, the newly proposed discriminative feature extraction block with context modeling scheme helps further improve the authentication performance against deepfake attacks, compared with three state-of-the-art methods on the GRID dataset.

6. REFERENCES

[1] Juergen Luettin, Neil A Thacker, and Steve W Beet, “Speaker identification by lipreading,” in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP’96*. IEEE, 1996, vol. 1, pp. 62–65.

[2] T Wark, D Thambiratnam, and S Sridharan, “Person authentication using lip information,” in *TENCON’97 Brisbane-Australia. Proceedings of IEEE TENCON’97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications (Cat. No. 97CH36162)*. IEEE, 1997, vol. 1, pp. 153–156.

[3] Chi Ho Chan, Budhaditya Goswami, Josef Kittler, and William Christmas, “Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 602–612, 2011.

[4] Jun-Yao Lai, Shi-Lin Wang, Alan Wee-Chung Liew, and Xing-Jian Shi, “Visual speaker identification and authentication by joint spatiotemporal sparse coding and hierarchical pooling,” *Information Sciences*, vol. 373, pp. 219–232, 2016.

[5] Feng Cheng, Shi-Lin Wang, and Alan Wee-Chung Liew, “Visual speaker authentication with random prompt texts by a dual-task cnn framework,” *Pattern Recognition*, vol. 83, pp. 340–352, 2018.

[6] Jiahui Sun, Shilin Wang, and Quanhai Zhang, “Visual speaker authentication by a cnn-based scheme with discriminative segment analysis,” in *International Conference on Neural Information Processing*. Springer, 2019, pp. 159–167.

[7] Chia-Wei Liao, Wei-Yang Lin, and Chia-Wen Lin, “Video-based person authetication with random passwords,” in *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 2008, pp. 581–584.

[8] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik, “Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,” *Applied Intelligence*, pp. 1–53, 2022.

[9] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia, “Deepfakes and beyond: A survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, 2020.

[10] Yisroel Mirsky and Wenke Lee, “The creation and detection of deepfakes: A survey,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[11] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar, “A lip sync expert is all you need for speech to lip generation in the wild,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 484–492.

[12] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, “First order motion model for image animation,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[13] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge, “Simswap: An efficient framework for high fidelity face swapping,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2003–2011.

[14] Pavel Korshunov and Sébastien Marcel, “Deepfakes: a new threat to face recognition? assessment and detection,” *arXiv preprint arXiv:1812.08685*, 2018.

[15] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu, “Generalizing face forgery detection with high-frequency features,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16317–16326.

[16] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic, “Lips don’t lie: A generalisable and robust approach to face forgery detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[17] Jiashang Hu, Shilin Wang, and Xiaoyong Li, “Improving the generalization ability of deepfake detection via disentangled representation learning,” in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3577–3581.

[18] Chen-Zhao Yang, Jun Ma, Shilin Wang, and Alan Wee-Chung Liew, “Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1841–1854, 2020.

[19] Jun Ma, Shilin Wang, Aixin Zhang, and Alan Wee-Chung Liew, “Feature extraction for visual speaker authentication against computer-generated video attacks,” in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1326–1330.

[20] H Hermansky, “Perceptual linear predictive (plp) analysis of speech,” *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 17–29, 1987.

[21] Davis E King, “Dlib-ml: A machine learning toolkit,” *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[22] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, “3d convolutional neural networks for human action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[25] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu, “Gcnet: Non-local networks meet squeeze-excitation networks and beyond,” in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[26] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu, “Pose-controllable talking face generation by implicitly modularized audio-visual representation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4176–4186.

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[29] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu 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.

[30] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu, “Multi-attentional deepfake detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

[31] Kaede Shiohara and Toshihiko Yamasaki, “Detecting deepfakes with self-blended images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18720–18729.