DEEPLIP: A BENCHMARK FOR DEEP LEARNING-BASED AUDIO-VISUAL LIP BIOMETRICS

Meng Liu[1], *Longbiao Wang*[1,*], *Kong Aik Lee*[2,*], *Hanyi Zhang*[1], *Chang Zeng*[3], *Jianwu Dang*[1]

[1]College of Intelligence and Computing, Tianjin University, China
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]National Institute of Informatics, Tokyo, Japan

2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 978-1-6654-3739-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASRU51503.2021.9688240

ABSTRACT

Audio-visual lip biometrics (AV-LB) has been an emerging biometrics technology that straddles auditory and visual speech processing. Previous works mainly focused on the front-end lip-based feature engineering combined with a shallow statistical back-end model. Over the past decade, convolutional neural network (CNN, or ConvNet) has been widely used and achieved good performance in computer vision and speech processing tasks. However, the lack of a sizeable public AV-LB database led to a stagnation in deep-learning exploration on AV-LB tasks. In addition to the dual audio-visual streams, one essential requirement on the video stream is the region of interest (ROI) around the lips has to be of sufficient resolution. To this end, we compile a moderate-size database using existing public databases. Using this database, we present a deep learning-based AV-LB benchmark, dubbed DeepLip[1], realized with convolutional video and audio unimodal modules, and a multimodal fusion module. Our experiments show that DeepLip outperforms the traditional lip-biometrics system in context modeling and achieves over 50% relative improvements compared with its unimodal system, with an equal error rate of 0.75% and 1.11% on the test datasets, respectively.

Index Terms: speaker recognition, audio-visual, lip biometrics, deep learning, visual speech

1. INTRODUCTION

Automatic speaker verification (ASV) systems play a crucial role in many applications, such as access control to e-commerce, teleworking and in-car systems. There is increasing concern that ASV systems are vulnerable to spoofing attacks, acoustically noisy environments, far-field detection, and other complex multifaceted scenarios. In this regard, audio-visual (AV) biometrics [1, 2, 3, 4] can be a viable solution. The incorporation of additional modalities alleviates the limitation of single modality and improve their joint performance. AV multimodal techniques have also achieved good performance on visual speech recognition (lipreading) [5], speech enhancement [6], speech separation [7] and emotion recognition [8].

Audio-visual lip biometrics (AV-LB) is a multimodal approach to speaker verification using both voice and lip dynamics [9]. Unlike speaker verification using face and voice [10, 11, 12], AV-LB focuses on the region-of-interest (ROI) around the lips during speaking. The mouth ROI is highly correlated to speech production since the lips, tongue, teeth and oral cavity are integral components of articulation [13]. In addition, evidence from lipreading research [14, 15] also indicates lip sequences reflect substantial speaker characteristics. AV-LB aims to extract correlated and complementary speaker characteristics from speech and lip dynamics.

Over the past decade, the rise of deep learning has significantly improved performances in video and speech processing tasks. However, these advancements have not been translated to AV-LB task. Traditional AV-LB pipelines employ delicate manual lip-based features combined with shallow statistical models, e.g., Gaussian mixture model (GMM) [16], hidden Markov model (HMM) [17] or support vector machine (SVM) [9]. For visual stream, conventional LB feature representation include appearance-based and shape-based features [1], which utilize lip geometry, parametric, or statistical models [18]. These features can be roughly divided into static feature processed with GMMs or dynamic features processed with HMMs. For audio stream, a universal background model (UBM) was trained and then mapped to target speakers [16].

On audio-based ASV, the time delay neural network (TDNN) based x-vector [19] and residual network (ResNet) based r-vector [20] architectures have achieved remarkable performance for extracting deep speaker embedding. Meanwhile, the progress in deep lipreading shows that 3D spatiotemporal convolution [21, 22] could better capture short-term lip dynamics compared to traditional lip-based feature engineering. For multimodal fusion, [1, 6, 23] discussed various approaches to realize the fusion of different modalities. In this paper, we look at a new form of speaker embedding combining both audio and visual modalities. To this end, we compile an AV-LB dataset as well as a deep-learning processing pipeline.

[1]https://github.com/DanielMengLiu/DeepLip

In this paper, our contribution focuses on the following aspects. Firstly, We establish an AV-LB database using public datasets. Secondly, we present a deep-learning based benchmark called DeepLip, which consists of a well-performed ConvNet for processing signals of speech spectrogram and lip image sequences. A standard deep-learning LB preprocessing pipeline is also introduced, doing away with laborious feature engineering and much prior expert knowledge. Thirdly, the effective fusion of auditory and visual speech embeddings shows that complementary and discriminative speaker information can be aggregated from speaker and lip dynamics.

2. AUDIO-VISUAL LIP BIOMETRICS

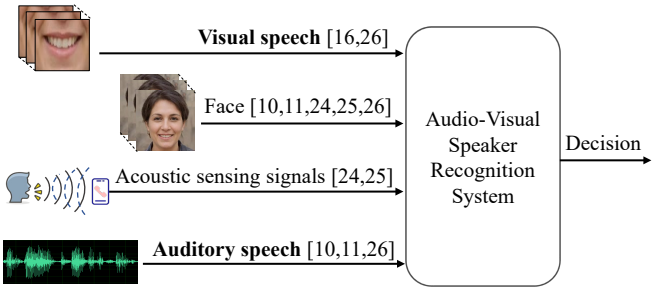

Fig. 1. Various input modalities of audio-visual biometric systems.

Audio-visual speaker recognition has attracted more and more attention recently. As shown in Figure. 1, existing AV-biometrics technologies mainly focuses on fusing facial features with auditory speech [10, 11], and those acquired with acoustic signals collected by mobile acoustic sensing [24, 25]. These methods have achieved promising improvements but have not explored the underlying dynamics between auditory speech and visual speech (lip motion). Lip biometrics aims to capture the physiological (i.e., static lip texture) and behavioral (i.e., dynamic lip movement) features, which contain substantial speaker characteristics. Therefore, lip biometrics is more than just recognizing the appearance of lip region [26]. In this paper, our proposed DeepLip leverages both the visual and auditory speech as the inputs of the ASV system.

Some advanced generative adversarial networks (GAN) and talking head models can achieve favorable performance in generating fake videos. However, realistic personalized lip movement is still beyond reach. The synthesized lip motions may well aligned with the lexical content, but are difficult to customize for each speaker. Lip motion are complicated. Even if different people say the same phrase, their lip motion would vary due to pronouncing habits and personal coarticulation effect.

3. THE DEEPLIP AV-LB SYSTEM

Human brain treats auditory and visual speech jointly. These findings have led to the consideration that speech perception works by extracting amodal information that takes the same form across modalities [27]. Figure 2 illustrates the overview of our convolutional AV-LB architecture, referred to as DeepLip. The architecture consists of two streams: an audio-only stream to process auditory speech, a visual-only stream to process the visual speech, and an audio-visual fusion module to fuse the audio and video speaker embeddings. As mentioned above, both the audio-only and visual-only stream are processed with ConvNet.

3.1. Visual embedding using multi-stage CNN

As shown in Figure 2, our visual stream uses a multi-stage convolutional neural network (MCNN) consisting of 3D, 2D and 1D convolutions to capture lip motion.

The CNN architecture of an AV-LB system differs from an audio-visual face recognition system mainly on **AV-LB requiring the fronted 3D convolution layer**. In face recognition, only 2D convolution is employed since the face sequences are usually downsampled to 2 frames every second or drops most of the frames. It is because there is no obvious transition between adjacent frames for face recognition and keeping those redundant frames will increase computation cost. In contrary, adjacent frames in a lip sequence may contain detailed lip dynamics and have coarticulation effect. Evidence from lipreading experiments also confirms that a mixture of 2D and 3D convolution [5, 21, 22] can extract more discriminative deep features than 2D structure alone.

Preprocessing should be done to extract lip sequences. Each video sequence from the dataset is processed by 1) performing face detection and face alignment (obtaining landmarks), 2) aligning each frame to a reference mean face shape, 3) cropping a fixed 96×96 pixels ROI from the aligned face image, 4) transforming the cropped image from RGB to gray level.

The visual system consists of an encoder ($\boldsymbol{F}_v$) followed by a nonlinear classifier ($\boldsymbol{C}_v$), which can be represented by

$$\boldsymbol{y}_v = \boldsymbol{C}_v(\boldsymbol{F}_v(\boldsymbol{x}_v)) \tag{1}$$

where $\boldsymbol{x}_v$ denotes the input of lip sequences and $\boldsymbol{y}_v$ is the final output of this visual network.

First, the grayscale lip sequences (a $B \times T \times H \times W$ tensor, corresponding to batch, frames, height and width) is segmented 29 frames every frame as CNN raw features, and a front-end 3D convolution with kernel size of $5 \times 7 \times 7$ to extract visual feature. It is followed by an 18-layer residual network (2D convolution) feature encoder. A global average pooling is applied to obtain visual speech embedding ($B \times C \times T$, where C is the output channel). Finally, a multiscale temporal convolutional network [28] is used to model

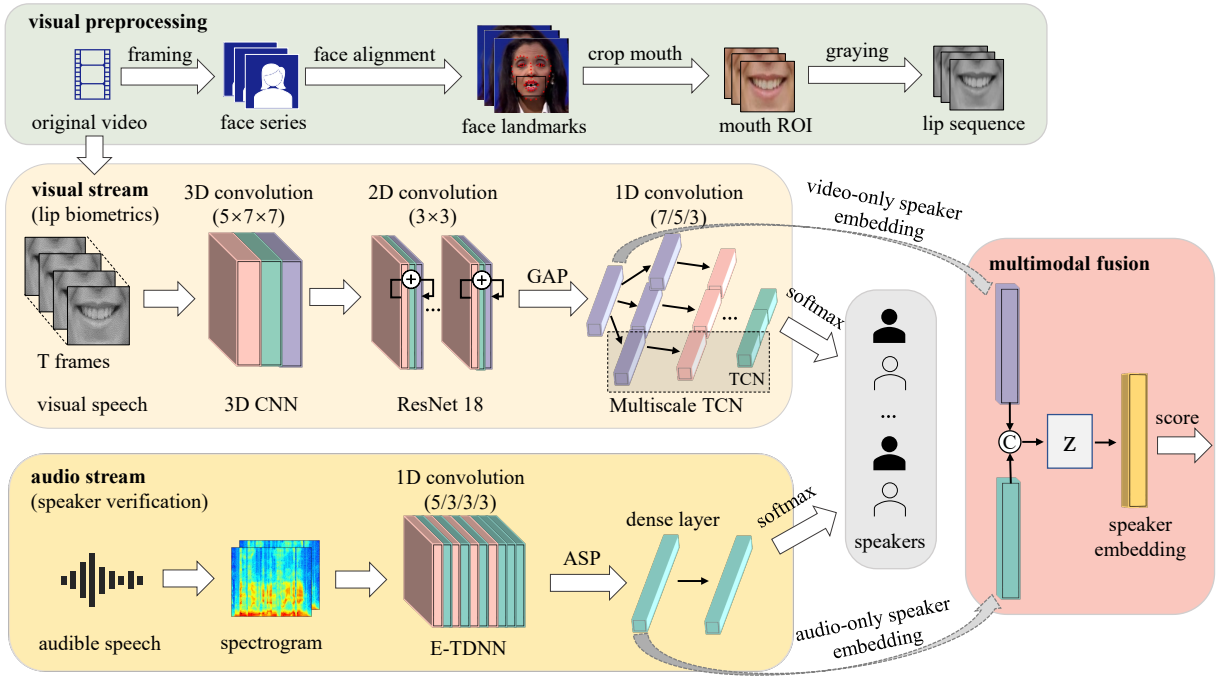

Fig. 2. DeepLip: an overview of our deep-learning based audio-visual lip biometrics architecture (GAP: global average pooling; ASP: attentive statistics pooling; TCN: temoral convolutional network).

the time-indexed sequences. Several basic temporal convolutional blocks are stacked sequentially to act as a deep feature sequence encoder. The temporal receptive field of a standard TCN is kept fixed for all activations at a specific layer [29]; multiple branches have been designed to provide variable receptive fields to fuse short-term and long-term temporal information during feature encoding. In the process, a variable-length augmentation procedure improves the generalization capabilities of the trained model when applied to sequences of varying lengths. 3D, 2D and 1D convolution play different roles in the visual stream during deep feature encoding. 3D convolution acts as a front-end feature extractor to extract fine-grained spatiotemporal lip dynamics. 2D convolution in a deep residual network can further process image information in every frame and extract deep embedding. 1D non-casual temporal convolution is explicitly designed to capture long-term temporal dynamic lip movements.

3.2. Audio embedding using x-vector

The x-vector based deep speaker embedding system has been widely recognized in large-scale speaker recognition tasks due to its stable and satisfactory performance. The extended time-delay neural network (E-TDNN) based x-vector architecture [30] improves the performance over the original x-vector system. Since speech is a time series, 1D convolution-based x-vector better captures long-term temporal dependencies of speech signals than 2D convolution-based r-vector [20]. Considering a balance between high performance and light weight, the x-vector system [30] (instead of better performed but heavier ECAPA-TDNN [31]), is a satisfactory framework for extracting audio-only speaker embedding. In addition, x-vector is more appropriate due to its light-weight structure. The initial frame layers in x-vector consist of 1-dimensional dilated convolutional layers interleaved with dense layers. Every filter has access to all the features of the previous layer or input layer. The task of the dilated convolutional layers is to build up the temporal context gradually. After the statistics pooling, two fully connected layers are utilized, with the first one acting as a bottleneck layer to generate the low-dimensional audio speaker characterizing embedding. The output layer computes posterior probabilities for the training speakers. Supposing $\boldsymbol{F}_a$ is the audio encoder which followed by a nonlinear classifier ($\boldsymbol{C}_a$), the final output $\boldsymbol{y}_a$ can be represented by

$$\boldsymbol{y}_a = \boldsymbol{C}_a(\boldsymbol{F}_a(\boldsymbol{x}_a)) \tag{2}$$

where $\boldsymbol{x}_a$ denotes the input of speech features.

3.3. Audio-visual fusion

Audio-visual fusion is of vital importance to prove the complementarity and relevance between the auditory speech and visual speech modalities. Early fusion, intermediate fusion and late fusion are three common approaches in multimodal integration [1, 6]. Intermediate fusion, late fusion and

Table 1. Database description of three databases used to compile the DeepLip database.

Database Acronym	Released Year	#Speakers (#Female)	#Utter -ances	Video Resolution, Frame Per Second	Content	Alternative Views
GRID	2006	34 (16 F)	34,000	360×288,25fps	controlled phrases	No
LombardGRID	2018	54 (30 F)	5,400	720×480,25fps	controlled phrases	Side view
TCD-TIMIT	2015	62 (30 F)	6,913	1920×1080,30fps	TIMIT sentences	30° view available

decision-level fusion (score fusion) are implemented in our DeepLip system, respectively.

For intermediate fusion and late fusion, the output embeddings of audio and visual stream are fed into a multimodal fusion module ($\boldsymbol{Z}$), shown as Figure 2. The final output ($\boldsymbol{y}_{av}$) of DeepLip is represented by

$$\boldsymbol{y}_{av} = \boldsymbol{Z}([\boldsymbol{F}_a(\boldsymbol{x}_a), \boldsymbol{F}_v(\boldsymbol{x}_v)]) \tag{3}$$

where audio-only embeddings and visual-only embeddings are concatenated as input of multimodal fusion module. $\boldsymbol{Z}$ is implemented with a linear encoder for intermediate fusion, and a one-dimensional identity matrix for late fusion. L1 normalization is applied before concatenating.

For decision-level fusion, final score of audio-visual system s_{av} is weighted by audio stream score s_a and visual stream score s_v:

$$s_{av} = \frac{d_a}{d_a + d_v} \cdot s_a + \frac{d_v}{d_a + d_v} \cdot s_v \tag{4}$$

where d_a and d_v denote the accuracy performance of audio-only and visual-only independent systems.

4. DEEPLIP DATABASE

The rapid development of deep-learning based multimodal face and speaker verification benefits from large-scale public database like VoxCeleb2. However, dataset constraints are the major obstacles to the study of deep-learning based AV-LB[2]. We overcome this by combining suitable open-source datasets. Among others we studied a number of audio-visual lipreading and biometrics datasets which include VoxCeleb2 [3], LRW-1000 [32], LRW [33], LRS, LRS2, LRS3 [34], GRID [35], LombardGRID [36] and TCD-TIMIT [37].

Our design principle is that an ideal AV-LB database satisfies the following conditions:

- Speaker labels. Most lipreading datasets ignore speaker labels, e.g., LRW, LRW-1000, LRS2 and LRS3.
- A sufficiently large number of speakers. Contrary to speech recognition, speaker recognition requires sufficient speakers instead of texts.
- High-quality recordings and high visual resolution to extract mouth region. VoxCeleb2 failed because there are many stage lectures.
- Continuous speech with good coverage of phonemes and visemes.
- Gender balanced and with a variety of facial hair (beard, moustaches, shaved) and skin tones.
- Available to other researchers.

Based on the above conditions, Table 1 describes the TCD-TIMIT, GRID and LombardGRID which we use to construct the DeepLip database. The GRID and LombardGRID corpora comprise phrases with fixed grammar, e.g., *bin blue at A one now*, while the TCD-TIMIT corpus contains phonetically rich sentences, e.g., *she had your dark suit in greasy wash water all year*. TCD-TIMIT and LombardGRID provides an alternative view of the visual speakers. Half of the LombardGRID data is designed for studying the lombard effect [36].

Table 2. Partitions of the DeepLip database.

Subset	Source	# Speakers	# Utterances
training	TCD-TIMIT	62 (30 F)	6,913
development	LombardGRID	18 (10 F)	1,774
test1	LombardGRID	36 (20 F)	3,541
test2	GRID	34 (16 F)	32,886

We checked and removed the erroneous videos with lost frames or faces. Table 2 shows the partitions of our DeepLip database[3]. The criteria behind the partitions is keeping original diversity (texts, recording environments and devices, etc.) cross sets as possible. We divide the original LombardGRID database into a development set with 18 speakers and a test1 set with 36 speakers. The development set is reserved for researchers to tune parameters or train a probabilistic linear discriminant analysis (PLDA) model. The development set and test2 set are non-overlapped, which could further validate whether the trained PLDA works well. Twenty thousand pairs of trials are randomly selected to form the test sets, including four thousands target and sixteen thousands nontarget pairs.

[2]VoxCeleb datasets cannot be used in AV-LB due to the low resolution of mouth region.

[3]https://github.com/ DanielMengLiu/DeepLip/database

Table 3. Performance comparison between traditional and proposed AV systems on two DeepLip test sets.

Modality	Model	Contrast	Test set	Measure	EER(%)	minDCF
audio-only	GMM-UBM	baseline [16]	LombardGRID	-	17.20	0.7688
audio-only	x-vector	ours	LombardGRID	PLDA	8.16	0.5524
audio-only	x-vector	ours	LombardGRID	Cos	12.63	0.5541
visual-only	GMM-UBM	baseline [16]	LombardGRID	-	21.03	0.9415
visual-only	MCNN	ours	LombardGRID	PLDA	2.70	0.1126
visual-only	MCNN	ours	LombardGRID	Cos	5.73	0.2089
audio-visual	GMM AV-LB	baseline [16]	LombardGRID	-	11.46	0.5895
audio-visual	DeepLip	ours	LombardGRID	Cos	**3.86**	**0.1744**
audio-only	GMM-UBM	baseline [16]	GRID	-	13.50	0.7558
audio-only	x-vector	ours	GRID	PLDA	14.01	0.8856
audio-only	x-vector	ours	GRID	Cos	9.62	0.6351
visual-only	GMM-UBM	baseline [16]	GRID	-	24.00	0.9762
visual-only	MCNN	ours	GRID	PLDA	7.13	0.3371
visual-only	MCNN	ours	GRID	Cos	8.65	0.3904
audio-visual	GMM AV-LB	baseline [16]	GRID	-	9.72	0.5994
audio-visual	DeepLip	ours	GRID	Cos	**5.25**	**0.2631**

5. EXPERIMENTS

5.1. Experimental setup

The training of each stream is performed independently. For video stream training, we train for 300 epochs with a cosine scheduler, an initial learning rate of 0.05 and a weight decay of 1e-4, using cross entropy and Adam as the loss function and optimizer. We use the random crop of 88×88 pixels and random horizontal flip for data augmentation. We split the lip sequences into several segments (every 29 frames corresponds to a segment, and segments less than 29 frames were abandoned). The basic TCN has four layers, and the multi-scale TCN is composed of three branches with convolutional kernel sizes of 3, 5 and 7.

For audio stream training, VoxCeleb1 and VoxCeleb2 datasets are pretrained for 30 epochs, and then we train for another 10 epochs on the training dataset. We employ SGD and AM-Softmax [38] as the optimizer and loss function, with a scale factor of 20, a margin of 0.25, an initial learning rate of 0.01, a weight decay of 1e-5 and a batch size of 256. The mel frequency cepstrum coefficient (MFCC) acts as the input speech feature, with the number of FFT points equal to 512 and the number of bins equal to 26. The configuration of E-TDNN is consistent with the classic setup described in [30]. Variable length augmentation is used in the training process. The universal background model with a mixture of 64 is trained on the training set, and then is mapped to target speakers. The PLDA model is trained on the development set with 150 principal components. Video and audio speaker embeddings are two 512 dimensional vectors.

5.2. Proposed DeepLip benchmark

A detailed performance comparison on two DeepLip test sets between traditional and our proposed systems is shown in Table 3. We train all of the single systems on training set only and evaluate these models on our test1 and test2 sets. Two approaches are employed to measure the distance between the two test embeddings: a non-parametric method cosine similarity and a parametric model PLDA trained on the development set. [16] reported the AV-LB performance of HMM-UBM (dynamic LB) and GMM-UBM (static LB). We choose GMM-UBM as our baseline, since previous results show no large performance gap between the two models, and we want the LB baseline system to be consistent with traditional speaker verification system (GMM-UBM).

The results is interesting. Firstly, there is no surprise that neural network systems significantly outperform traditional statistical models with adequate training data (dozens of speakers). This further reveals the necessity and urgency of promoting deep-learning researches in AV-LB. Secondly, it demonstrates that traditional audio-only GMM-UBM system performs better than visual-only GMM-UBM system while deep-learning based audio-only x-vector system performs worse than visual-only MCNN system. This reveals that the visual speech (visual lip motion) contains satisfactory potential speaker characteristics, since the visual-only GMM-UBM baseline captures static lip features while our MCNN can capture shor-term and long-term dynamic lip motion with 3D and 1D convolution. Thirdly, PLDA does not seem to have a stable performance. PLDA significantly improves the performance on test1 set in both the video and

audio streams. The improvement decreases for the test2 data of the video-only stream and is even worse for the test2 data of the audio-only stream. This is mainly because the development and test1 data are all from LombardGRID dataset. It seems that PLDA overfits for the LombardGRID dataset, so we use cosine similarity as a measurement method in our audio-visual system. Fourthly, our proposed audio-visual DeepLip benchmark outperforms traditional GMM AV-LB system, with equal error rates (EERs) of 3.86% and 5.25% on the LombardGRID and GRID test sets, respectively. Compared with our best unimodal systems (here visual-only), our DeepLip benchmark can achieve over 40% relative error reduction.

5.3. Deep-learning based AV-LB with transfer learning

Table 4. Performance comparison among audio-only, video-only and audio-visual systems using cosine similarity measurement.

Modality	Test set	Fusion	EER(%)
audio-only	LombardGRID	single system	**1.98**
audio-only	GRID	single system	**2.12**
video-only	LombardGrid	single system	5.73
video-only	GRID	single system	8.65
audio-visual	LombardGrid	intermediate	1.21
audio-visual	GRID	intermediate	1.58
audio-visual	LombardGrid	late	0.84
audio-visual	GRID	late	**1.11**
audio-visual	LombardGrid	decision	**0.75**
audio-visual	GRID	decision	1.16

To the best of our knowledge, training a neural network with dozens of speakers may not be fully discriminative. With transfer learning technology in deep learning, pretraining on public large-scale speaker verification database and finetuning on our training set may improve the audio-stream and audio-visual performance. Therefore, we enhance the audio-only system via transfer learning. To this end, VoxCeleb audio data and the training sets are used to pretrain and fine tune, respectively. Table 4 shows a performance comparison among audio-only, video-only and audio-visual systems using cosine similarity measurement. The enhanced audio-only system can achieve an EER of 1.98% on LombardGRID test set and 2.12% on GRID test set, which serves as our best unimodal system (here audio-only).

Late fusion, which concatenates the audio-only and video-only embeddings, significantly improves the performance of speaker recognition, with EERs of 0.84% and 1.11%. Remarkable complementary speaker information existed between visual speech and speech, revealing the potential to find a good feature fusion method. The score fusion for the audio-only and video-only systems also achieve a relative improvement of approximately 50%. Collection of large-scale (over thousands) visual speakers may be difficult, but easy for audio speakers with existing public databases. The results inspire us that using DeepLip system additional modality can have around 40% ~ 50% improvement compared with the best unimodal system, no matter the best unimodal system is audio-only or visual-only system. This interesting finding may significantly improve performance of current person authentication system with a low cost of data collection. Above experimental results show that AV-LB can be one of the potential approaches in future speaker/person recognition.

6. CONCLUSIONS AND FUTURE WORK

Audio-visual lip biometrics is an emerging multimodal speaker recognition technique. Unlike multimodal face and speaker recognition system, it operates on specific lip sequences of smaller image size but deeper depth. To fill the vacuum of deep-learning methods and promote the development of traditional AV-LB, we have organized the DeepLip database with easily available public datasets. We also presented a deep-learning based AV-LB benchmark to leverage deep audio and visual speaker embeddings as well as their complementarity information. Experimental results show the effectiveness of deep learning methods compared with the conventional AV-LB system. Furthermore, with transfer learning on large-scale audio speaker verification, we obtained performance improvement over 50% compared with that of our best unimodal system through audio-visual fusion, which implies deep-learning based AV-LB systems may be promising in future.

With the convenience of audio-visual data collection using mobile devices, we predict that audio-visual lip biometrics may soon become a popular research direction. More deep-learning techniques are expected to be introduced. This work serves as a benchmark and proves the feasibility technically. However, It still cannot focus on the correlation and alignment (multimodal integration) cross modalities due to the limitation of the texts; this correlation will be explored in our future research.

7. ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant 61771333 and the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

8. REFERENCES

[1] Petar S Aleksic and Aggelos K Katsaggelos, "Audio-visual biometrics," *Proceedings of the IEEE*, vol. 94,

no. 11, pp. 2025–2044, 2006.

[2] Seyed Omid Sadjadi, Craig S Greenberg, Elliot Singer, Douglas A Reynolds, Lisa Mason, and Jaime Hernandez-Cordero, "The 2019 nist audio-visual speaker recognition evaluation," *Proc. Speaker Odyssey (submitted), Tokyo, Japan*, 2020.

[3] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[4] 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.

[5] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[6] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[7] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–11, 2018.

[8] Ruo-Hong Huan, Jia Shu, Sheng-Lin Bao, Rong-Hua Liang, Peng Chen, and Kai-Kai Chi, "Video multimodal emotion recognition based on bi-gru and attention fusion," *Multimedia Tools and Applications*, pp. 1–28, 2020.

[9] Maycel Isaac Faraj and Josef Bigun, "Speaker and speech recognition by audio-visual lip biometrics," in *The 2nd International Conference on Biometrics, Seoul Korea*. Citeseer, 2007.

[10] Ruijie Tao, Rohan Kumar Das, and Haizhou Li, "Audio-visual speaker recognition with a cross-modal discriminative network," *Proc. Interspeech 2020*, pp. 2242–2246, 2020.

[11] Yanmin Qian, Zhengyang Chen, and Shuai Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[12] Leda Sarı, Kritika Singh, Jiatong Zhou, Lorenzo Torresani, Nayan Singhal, and Yatharth Saraf, "A multi-view approach to audio-visual speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6194–6198.

[13] Hans-Heinrich Bothe and Frauke Rieger, "Visual speech and coarticulation effects," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1993, vol. 5, pp. 634–637.

[14] Michael Wand and Jürgen Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," *Proc. Interspeech 2017*, pp. 3662–3666, 2017.

[15] Ibrahim Almajai, Stephen Cox, Richard Harvey, and Yuxuan Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2722–2726.

[16] Shi-Lin Wang and Alan Wee-Chung Liew, "Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power," *Pattern Recognition*, vol. 45, no. 9, pp. 3328–3335, 2012.

[17] 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.

[18] Enrique Gómez, Carlos M Travieso, Juan C Briceño, and Miguel A Ferrer, "Biometric identification system by lip shape," in *Proceedings. 36th Annual 2002 International Carnahan Conference on Security Technology*. IEEE, 2002, pp. 39–42.

[19] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Interspeech*, 2017, pp. 999–1003.

[20] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[21] Themos Stafylakis and Georgios Tzimiropoulos, "Deep word embeddings for visual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4974–4978.

[22] Xuejuan Chen, Jixiang Du, and Hongbo Zhang, "Lipreading with densenet and resbi-lstm," *Signal, Image and Video Processing*, vol. 14, no. 5, pp. 981–989, 2020.

[23] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[24] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Linghe Kong, and Minglu Li, "Lip reading-based user authentication through acoustic sensing on smartphones," *IEEE/ACM Transactions on Networking*, vol. 27, no. 1, pp. 447–460, 2019.

[25] Libing Wu, Jingxiao Yang, Man Zhou, Yanjiao Chen, and Qian Wang, "Lvid: A multimodal biometrics authentication system on smartphones," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1572–1585, 2019.

[26] Xin Liu and Yiu-ming Cheung, "Learning multi-boosted hmms for lip-password based speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 2, pp. 233–246, 2013.

[27] Lawrence D Rosenblum, "Speech perception as a multimodal phenomenon," *Current Directions in Psychological Science*, vol. 17, no. 6, pp. 405–409, 2008.

[28] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6319–6323.

[29] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic, "Towards practical lipreading with distilled and efficient models," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7608–7612.

[30] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

[31] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. Interspeech 2020*, pp. 3830–3834, 2020.

[32] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin 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.

[33] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 87–103.

[34] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[35] 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.

[36] Najwa Alghamdi, Steve Maddock, Ricard Marxer, Jon Barker, and Guy J Brown, "A corpus of audio-visual lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[37] Naomi Harte and Eoin Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[38] Yi Liu, Liang He, and Jia Liu, "Large margin softmax loss for speaker verification," *Proc. Interspeech 2019*, pp. 2873–2877, 2019.