Lip Reading using Facial Expression Features

Tatsuya Shirakata and Takeshi Saitoh
Kyushu Institute of Technology
680–4 Kawazu, Iizuka, Fukuoka, 820–8502, Japan
Email: saitoh@ces.kyutech.ac.jp

***Abstract*—Lip-reading technology that estimates speech content only from visual data without using audio data is expected as a next-generation interface. It can be used even in a situation where audio recognition is difficult such as in a high noise environment or a person with a speech disorder. The conventional methods use only features around the lips. This paper proposes a method to integrate facial expression features; Expression-based feature and action unit-based feature into the lip-reading method. Evaluation experiments are conducted with three public databases of OuluVS, CUAVE, and CENSREC-1-AV. As a result, it is confirmed that the recognition accuracy is improved by integrating the facial expression feature for all databases.**

I. Introduction

Lip-reading technology uses lip movements that can be acquired simultaneously as speech communication that is most familiar to humans. Furthermore, this is a technique for recognizing utterance contents using only visual data without audio data. It is expected to be used in cases where the use of speech recognition technology is difficult, such as in a high-noise environment where it is challenging to obtain audio data, in a public place where it is challenging to produce a voice, and for communication support for persons with disabilities who cannot speak due to laryngectomy.

Although lip-reading was studied several decades ago, yet, there are many challenges in lip reading tasks, such as the recognition target: single sound, isolated word, continuous word, and sentence, available modalities: audio-visual or only-visual, face direction, and language. It has not been put to practical use. Related researches are briefly introduced in Section II.

Observing the face during the speech, it is evident that the lips move with the speech, but the skin around the lips also moves. Usually, although the lip-reading task uses speech scenes without expression, the expression of the speaker is not expressionless but changes according to the content of the conversation in the actual conversation. However, no facial expression information is used in the conventional lip-reading method. This paper proposes a new lip-reading method that integrates facial expression features.

Our contribution is to improve the lip-reading method by using facial expression features, rather than proposing a lip-reading method superior to the state-of-the-art. This paper shows evaluation experiments with three public databases and confirmed that the recognition accuracy is improved by using the facial expression.

The rest of this paper is organized as follows: Related researches are introduced in Section II. Section III describes the proposed method. Section IV introduces databases used in the experiment and describes experimental results. This paper concludes in Section V.

II. Related Research

Various methods have been proposed in the lip-reading field. In the conventional approaches, various hand-craft features extraction methods are proposed. There are roughly four approaches for feature extraction [1]: (1) appearance-based [2], [3], (2) motion-based, [4] (3) geometric-feature-based, and (4) model-based [5]. These features are fed to the Hidden Markov Model (HMM) [5].

Recently, a deep learning-based approach has been applied to learn features from either audio-visual or visual data for the tasks of lip reading, and the accuracy is significantly improved. Noda et al. [6] apply a convolutional neural network (CNN) as the visual feature extraction mechanism. These features are fed to HMM with Gaussian mixtures for the task of recognizing isolated words. Saitoh et al. [7] proposed a concatenated frame image (CFI) and evaluated it with a public database OuluVS2. Chung and Zisserman proposed several CNN architectures and demonstrated that the proposed multiple tower model obtained the highest performance. They also created several datasets of The Oxford-BBC Lip Reading in the Wild (LRW) [8], and The Oxford-BBC Lip Reading Sentences 2 (LRS2) [9]. Mesbah et al. [10] proposed Hahn CNN, which uses Hahn moments. The Hahn moments filters are embedded before the first layer of CNN. They evaluated their method with AVLetter, OuluVS2, and LRW. Petridis et al. [11] proposed an end-to-end method based on fully-connected layers and LSTM for the lip-reading task. They reported the comparison experiments on OuluVS2, CUAVE, AVLetters, and AVLetters2.

III. Proposed Method

A. Overview

Figure 2 shows an overview of our proposed approach. Our approach is based on the existing method [12]. This is a method that integrates two types of hand-craft features: appearance-based and motion-based, and this structure makes it easy to integrate other features. For the above reasons, this paper adopts this method as the baseline method.

The face detector is first applied to the input face image and extracts a face ROI, as shown in a red rectangle of the top face image of Fig. 1. Next, the facial feature points detection process [13] is applied, and 68 facial points of

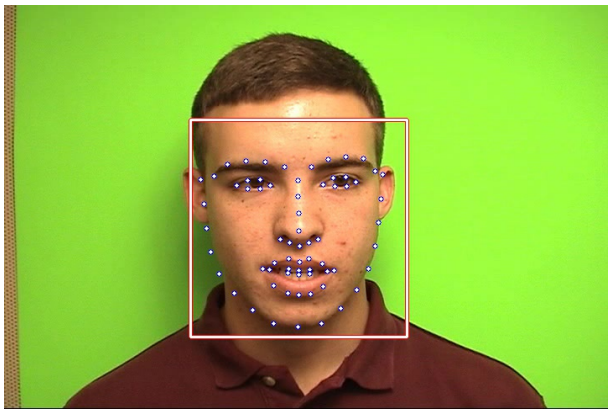

Fig. 1. Face and feature points detection.

$P_i, i = 0, 1, \cdots, 67$ are detected. The green points of the top face image of Fig. 2 are detected points.

Not all speech scenes have the same recording environment. For this reason, the face position and size differ depending on the speech scene. This difference is undesirable, and the face alignment process is applied using detected facial points. Two values of $d_{eye} = |P_{36} - P_{45}|$ and $\theta = \angle P_{36}P_{45}$ are calculated. P_{36} and P_{45} are the feature point coordinates outside the left eye and the right eye, respectively. d_{eye} is a distance between two eyes, and θ is an angle between two eyes. Next, an affine transformation is applied using d_{eye} and θ, Specifically, the scale is changed so that d_{eye} becomes 200 pixels, and the image is rotated so that θ becomes 0 degree.

The following feature extraction is applied using the detected face image and the feature points. Recognition processing using Gated recurrent unit (GRU) is applied with the obtained feature values as input data. There are several fusion schemes, early fusion, and late fusion. The proposed approach uses separate GRU flows the different feature values and apply late fusion.

B. Features

This section describes six feature extraction methods.

1) Appearance-based Feature (AF): Generally, the appearance-based feature is calculated by extracting the ROI around the lip. This feature can contain not only lip appearance but also information on teeth and tongue inside the lip. Therefore, this is a useful feature for the lip-reading task.

Before deep learning become popular, most of the appearance-based features use a gray-scale image, and this is either used as a feature vector directly or applying some image processing, such as Principal Component Analysis (PCA) and Discrete Cosine Transform (DCT) [2], [3]. After deep learning became popular, auto-encoder has been used to calculate typical appearance-based features [12].

The auto-encoder neural network is an unsupervised learning algorithm that applies backpropagation, setting the target values $\hat{x}$ to be equal to the inputs x. This network is data-specific, lossy, and learned automatically from data examples. It is one of the data compression algorithms.

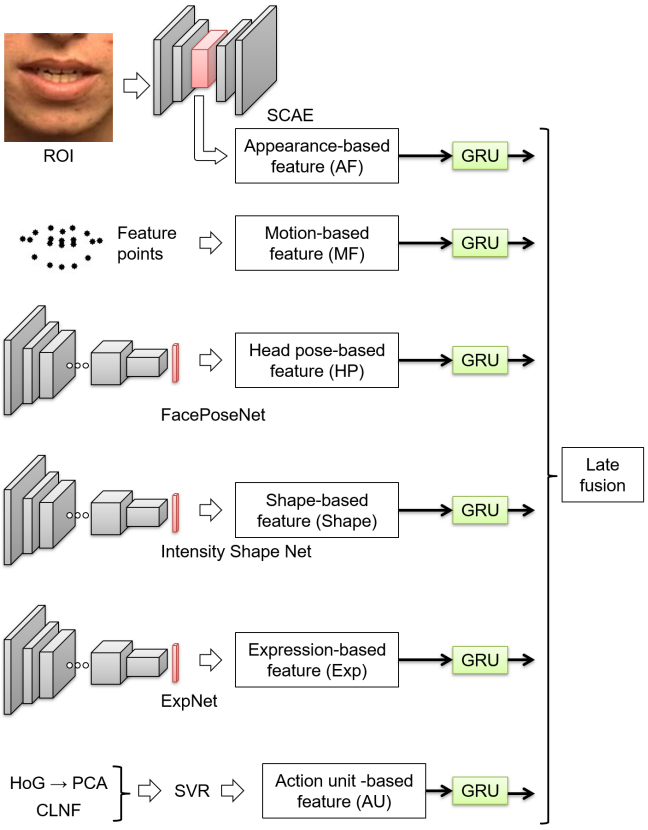

Fig. 2. Various features and fusion approach.

Iwasaki et al. [12] used stacked convolution auto-encoder (SCAE) [14]. In [12], the size of input image is 64×64 [pixels]. The encoding process consists of three pairs of convolutional layers, the pooling layer, and the decode process to decompress the image after wrapping. In this paper, 256 values of the bottleneck layer are used as the appearance-based feature.

2) Motion-based Feature (MF): The appearance-based feature is easily influenced by the color difference based on gender, race, and lighting condition. Thus, several motion-based features were proposed by [4], [12]. Shiraishi and Saitoh proposed the motion-based feature using optical flow [4]. However, the optical flow is energetically sensitive to head motion, even though it has little effect on lip-reading tasks. Thus, Iwasaki et al. [12] proposed the motion-based feature based on the facial feature points.

A subtraction value between current frame and next frame at each feature point is defined as a motion-based feature by $d_*(i, f) = P_*(i, f) - P_*(i, f + 1)$, where i is a feature point number, f means a frame number, and $P_*(i, f)$ is a coordinate of i-th feature point. A symbol $*$ is either x or y. This feature is not a distance value; it has negative or positive value.

An important point to be considered in the motion-based

feature is which feature points are used. On this problem, Saitoh investigated an efficient face model for lip-reading task [5], and described that eyes and eyebrows are not useful for lip reading. Thus, based on this fact, this paper uses only 20 feature points of the lips.

3) Head pose-based Feature (HP): The head pose changes independently of the movement of the lips. However, many people speak while moving their heads. It is presumed that the head movement changes depending on the content of the conversation and the conversation partner. The head pose is not considered in conventional lip-reading methods. This paper defines the head pose as a feature and investigate its effectiveness.

Various head pose estimation methods have been proposed. Chang et al. [15] proposed a FacePoseNet (FPN) that is a CNN-based six degrees of freedom (6DoF) face pose estimation method directly from image intensities. FPN uses AlexNet architecture [16] with its initialized weights provided by [17]. Their method defines 6DoF 3D face pose $\mathbf{h} = (r_x, r_y, r_z, t_x, t_y, t_z)$, where (r_x, r_y, r_z) are Euler angles (pitch, yaw, and roll), and $(t_x, t_y, t_z)^T$ is three translations of 3D head pose. This paper uses FPN that is implemented by ExpNet and obtains six values as HP-based features.

4) Shape-based Feature (Shape): Trans et al. [18] proposed CNN-based 3D morphable face models (3DMM), which regress 3DMM shape and texture parameters directly from an input image. They use the very deep ResNet architecture [19] with 101 layers.

The 3DMM is a feature based on three dimensions, not two dimensions, like the appearance-based feature. Thus, this paper uses Trans's model and obtains 99 values as shape-based features.

5) Expression-based Feature (Exp): In general, facial expression estimation is treated as a classification problem, and classifies images or videos into discrete sets of facial expression classes, such as anger, disgust, fear, happiness, sadness, and surprise [20] or action units [21]. On the other hand, Chang et al. [22] study the expression regression problem.

In [22], a deep neural network ExpNet which uses ResNet101 architecture [19] was proposed. ExpNet is applied directly to the face image's intensities and regresses a 29D vector of 3D expression coefficients. This method is the first approach to directly estimate 3D expression coefficients without requiring or involving facial landmark detection. 29D values cannot be easily labeled manually by human operators and have no natural interpretation. This paper uses 29 values as expression-based features.

6) Action unit-based Feature (AU): Facial Action Coding System (FACS) is a system to taxonomize human facial movements by their appearance on the face. FACS encodes movements of individual facial muscles from slightly different instant changes in facial appearance. Using FACS, it is possible to code nearly any anatomically possible facial expression, deconstructing it into the specific Action Units (AU) that produced the expression. It is a common standard to describe facial expressions objectively.

Baltrusaitis et al. [21] propose a real-time intensity value estimation method of facial action unit, that is not a deep learning-based a hand-craft based approach. They extract two types of features after face alignment is applied. The first one is an appearance-based feature obtained by applying Principal Component Analysis (PCA) to Histograms of Oriented Gradients (HOGs) features. The second one is a geometry-based feature obtained by the non-rigid shape parameters and Constrained Local Neural Field (CLNF) parameters. Support Vector Regression (SVR) is used for AU intensity estimation. They used three datasets: DISFA, BP4D-Spontaneous, and SEMAINE, and applied the cross-dataset training. In their method, 17 AUs are estimated, and this paper uses all 17 AUs as AU-based features.

C. Recognition by GRU

Recurrent Neural Networks (RNN) is a class of artificial neural network and belong to the most promising algorithms out there at the moment due to their internal memory capability. RNN allows us to exhibit temporal dynamic behavior for a time sequence. Unlike feedforward neural networks, RNN can use their internal state (memory) to process input sequences. It makes them ideal for applications such as speech recognition and time-series forecasting. Gated Recurrent Unit (GRU) aims to solve the vanishing gradient problem, which comes with a standard RNN [23]. GRU can also be considered as a variation on the Long Short-Term Memory (LSTM) because both are designed similarly and, in some cases, produce equally excellent results.

In this paper, three kinds of LSTMs (LSTM original [24], LSTM forget [25], LSTM peep [26]) and GRU were used as a preliminary experiment, and confirmed that GRU can obtain high recognition accuracy. Therefore, GRU is adopted as a recognition method.

D. Fusion Scheme

As described above in III-A, this paper uses late fusion. The early fusion was also carried out in a preliminary experiment. However, the late fusion obtained higher recognition accuracy than the early fusion under all the conditions described in the experiment in which multiple features were integrated. For this reason, this paper adopts a late fusion scheme.

IV. Evaluation Experiments

A. Databases for Lip-reading

There are few public databases in the lip-reading research field compared to the speech recognition research field. The public databases used in the lip-reading tasks are as follows: CUAVE [27], Grid [28], OuluVS [29], OuluVS2 [30], LRW [8], and LRS [9]. These are English speech scenes. CENSREC-1-AV [31] and SSSD [32] are Japanese speech scene. This section briefly introduces three databases used in the experiments in this paper. All speakers speak without expression in all databases.

TABLE I
THREE DATABASES USED IN THE EXPERIMENT.

name	CUAVE [27]	OuluVS [29]	CENSREC-1-AV [31]
contents	10 digits	10 phrases	10 digits
language	English	English	Japanese
number of speakers	36	20	93
image size [pixel]	720×480	720×576	720×480
frame rate [fps]	29.97	25	29.97
total image number	1,790	997	1,350

Fig. 3. Sample images of CUAVE (left: S01, right: S02).

1) CUAVE [27]: CUAVE (The Clemson University Audio-Visual Experiments) database contains speakers talking in a frontal and profile poses. This consists of 36 speakers; 19 males and 17 females. Each speaker uttered both isolated and continuous ten digits ("zero", "one", "two", "three", "four", "five", "six", "seven", "eight", and "nine") in English. Each isolated digit sequence is broken into the four tasks. This paper used the speech scene of the first normal task, where each speaker speaks 50 digits while standing still naturally. The size of image is 720×480 [pixel] and its frame rate is 29.97fps. The background of each speaker is green, as shown in Fig. 3.

2) OuluVS [29]: OuluVS is a database constructed by the research group of Oulu University. This consists of 20 subjects; 17 males and 3 females, uttering ten daily-use short English phrases ("excuse me", "goodbye", "hello", "how are you", "nice to meet you", "see you", "I am sorry", "thank you", "have a good time", and "you are welcome"). Each subject utters these phrases five times in front of the camera. The size of image is 720×576 [pixel] and its frame rate is 25fps. The background of each speaker is white, as shown in Fig. 4.

Fig. 4. Sample images of OuluVS (left: P002, right: P003).

Fig. 5. Sample ROI images of CENSREC-1-AV (left: FBJ, right: MBE).

3) CENSREC-1-AV [31]: CENSREC-1-AV is a Japanese audio-visual speech corpus. In the original database, the speaker speaks 1 to 7 Japanese digits. Each digit is pronounced as /i-chi/(one), /ni/ (two), /sa-N/ (three), /yo-N/ (four), /go/ (five), /ro-ku/ (six), /na-na/ (seven), /ha-chi/ (eight), /kyu/ (nine), /ze-ro/ (zero) or /ma-ru/ (zero). There is 3,234 training data and 1,963 test data. Only ROI image around the lip as shown in Fig. 5 is released to the public. The frame rate is 29.97fps. The speech scenes were taken from the front, and the background color was almost blue.

This experiment focuses on an isolated word recognition task, and our method requires the whole face image. Thus, we got the original whole face image from the author of [31]. The size of the image is 720×480 [pixels]. We selected samples in which only one digit was uttered.

B. Experimental Conditions

In this experiment, we used three databases of CUAVE, OuluVS, and CENSREC-1-AV, introduced in Section IV-A, for evaluating the recognition accuracy by our method.

There are some test protocols for the lip-reading task: speaker-dependent (SD), speaker semi-dependent (SD), and speaker-independent (SI). The most challenging task is SI; then, we tested the recognition task on SI. As for the recognition experiments of CUAVE and OuluVS, the leave-one-person-out evaluation was applied. That is, we trained the GRU model by all the speakers in the database except one and tested it by one speaker who has not used it in training. We repeated this task for all speakers in the database. We computed the averaging accuracy from the individual testing rounds. On the other hand, as for the recognition experiments of CENSREC-1-AV, the training data and test data are defined in [31]. Then, the hold-out evaluation was applied.

In this experiment, dlib[1] was used for face detection and feature points detection described in Section III. Three kinds of features; HP, Shape, and Exp, were extracted by applying ExpNet[2]. As for AU, we used OpenFace2.0[3] [33]. We used Keras deep learning library with the TensorFlow backend for training and tested the GRU model.

In this experiment, we calculate the six types of features described in Section III-B. Here, 11 recognition experiments shown in Table II were conducted by combining six features.

C. Recognition Results

The experimental results can be found in Table II. In this table, the first six columns are the features used in each experi-

[1]http://dlib.net/
[2]https://github.com/fengju514/Expression-Net
[3]https://github.com/TadasBaltrusaitis/OpenFace

ment, and the last three columns show the average recognition accuracy for each dataset. The number in parentheses of the feature indicates the number of dimensions.

The above six conditions are when each feature is used alone. The recognition accuracies of HP and Shape, where lip movement is not included in the feature values, are low. High recognition accuracies are obtained using MF and AF, which are the existing methods for lip reading. The recognition accuracy of Condition 7, which is the conventional method, is high. However, the proposed method of conditions 10 and 11, including the facial expression features, achieved higher recognition accuracies than the conventional method. MF obtains the movement of feature points around the lip, and AF obtains the lip and the skin around it. On the other hand, the facial expression feature implicitly includes a facial part and skin movement, and it complements MF and AF. Therefore, it is considered that the accuracy is improved.

Comparing the three databases, OuluVS obtained higher recognition accuracy than the other two databases. All databases have ten classes. OuluVS is speech scenes of short phrases in daily conversation, while CUAVE and CENSREC-1-AV are speech scenes of digits. The number of sounds in the digit speech scene is small, and the mouth movement is monotonous, while the phrase speech scenes are more complex than digit speech scenes. From Table I, the frame rate of CUAVE and CENSREC-1-AV and the frame rate of OuluVS are different. However, since all databases have different utterance lengths and languages, it is presumed that the frame rate has little effect.

V. Conclusion

While the facial expression feature was not used in the conventional method for lip reading, this paper proposed a new method using facial expression features. This paper evaluated the effectiveness of the proposed method with three public databases. As a result, it was confirmed that the recognition accuracy was improved by using the facial expression feature. Moreover, it was also shown experimentally that head pose and shape features were not useful for the lip-reading task.

Although the recognition accuracy was improved by the proposed method, it is not enough yet. Future challenges include further improvement in recognition accuracy. It is also a future task to evaluate the proposed method for speech scenes with actual facial expressions. In this paper, three databases were used. We have not fully discussed the differences in language and frame rate. These are also future tasks.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 16H03211, 17H01840, and 19KT0029.

References

[1] Z. Zhou, G. Zhao, X. Hong, and M. Pietikainen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, pp. 590–605, 2014.

[2] C. Bregler and Y. Konig, ""eigenlips" for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1994, pp. 669–672.

[3] P. J. Lucey, G. Potamianos, and S. Sridharan, "Patch-based analysis of visual speech from multiple views," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2008, pp. 69–73.

[4] J. Shiraishi and T. Saitoh, "Optical flow based lip reading using non rectangular ROI and head motion reduction," in *11th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[5] T. Saitoh, "Efficient face model for lip reading," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2013, pp. 227–232.

[6] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," in *INTERSPEECH*, 2014, pp. 1149–1153.

[7] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikainen, "Concatenated frame image based cnn for visual speech recognition," in *ACCV 2016 Workshops, LNCS 10117*, 2017, pp. 277–289.

[8] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision (ACCV)*, 2016.

[9] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6447–6456.

[10] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa, and M. Daoudi, "Lip reading with hahn convolutional neural networks," *Image and Vision Computing*, vol. 88, pp. 76–83, 2019.

[11] S. Petridis, Y. Wang, P. Ma, Z. Li, and M. Pantic, "End-to-end visual speech recognition for small-scale datasets," *Pattern Recognition Letters*, vol. 131, pp. 421–427, 2020.

[12] M. Iwasaki, M. Kubokawa, and T. Saitoh, "Two features combination with gated recurrent unit for visual speech recognition," in *IAPR Conference on Machine Vision Applications (MVA)*, 2017, pp. 300–303.

[13] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[14] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *21th International Conference on Artificial Neural Networks*, 2011, pp. 52–59.

[15] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Faceposenet: Making a case for landmark-free face alignment," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[17] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4838–4846.

[18] A. T. Trân, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3D morphable models with a very deep neural network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5163–5172.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv:1804.08348*, 2018.

[21] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *11th IEEE Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[22] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "ExpNet: Landmark-free, deep, 3D facial expressions," in *13th IEEE Conference on Automatic Face and Gesture Recognition (FG)*, 2018.

[23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] F. A. Gers, J. A. Schmidhuber, and F. A. Cummins, "Learning to forget: Continual prediction with lstm," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[26] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of Machine Learning Research*, vol. 3, pp. 115–143, 2003.

TABLE II
RECOGNITION RESULTS.

	features						databases		
condition	MF (40)	AF (256)	HP (6)	Shape (99)	Exp (29)	AU (17)	OuluVS [%]	CUAVE [%]	CENSREC-1-AV [%]
(1) [12]	○	—	—	—	—	—	73.2	76.4	54.3
(2) [12]	—	○	—	—	—	—	79.1	74.9	72.5
(3)	—	—	○	—	—	—	13.4	19.9	10.0
(4)	—	—	—	○	—	—	13.7	27.3	13.1
(5)	—	—	—	—	○	—	67.4	67.4	23.7
(6)	—	—	—	—	—	○	69.8	71.2	59.2
(7) [12]	○	○	—	—	—	—	83.1	79.9	74.3
(8)	○	○	○	—	—	—	79.4	75.5	24.0
(9)	○	○	—	○	—	—	79.3	75.1	74.5
(10)	○	○	—	—	○	—	85.6	83.1	74.5
(11)	○	○	—	—	—	○	**86.6**	**83.4**	**77.1**

[27] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker, speaker-independent feature study, and baseline results using the cuave multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 1189–1201, 2002.

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

[29] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[30] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "OuluVS2: a multi-view audiovisual database for non-rigid mouth motion analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[31] S. Tamura, C. Miyajima, N. Kitaoka, T. Yamada, S. Tsuge, T. Takiguchi, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2010.

[32] T. Saitoh and M. Kubokawa, "SSSD: Speech scene database by smart device for visual speech recognition," in *24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3228–3232.

[33] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *13th IEEE Conference on Automatic Face and Gesture Recognition (FG)*, 2018.