future internet

MDPI

Article

Multi-Angle Lipreading with Angle Classification-Based Feature Extraction and Its Application to Audio-Visual Speech Recognition †

Shinnosuke Isobe [1,*], Satoshi Tamura [2], Satoru Hayamizu [2], Yuuto Gotoh [3] and Masaki Nose [3]

[1] Graduate School of Natural Science and Technology, Gifu University, 1-1 Yanagido, Gifu 501-1193, Japan
[2] Faculty of Engineering, Gifu University, 1-1 Yanagido, Gifu 501-1193, Japan; tamura@info.gifu-u.ac.jp (S.T.); hayamizu@gifu-u.ac.jp (S.H.)
[3] Ricoh Company, Ltd., 2-7-1 Izumi, Ebina 243-0460, Kanagawa, Japan; yuuto.gotoh@jp.ricoh.com (Y.G.); masaki.nose@jp.ricoh.com (M.N.)
* Correspondence: isobe@asr.info.gifu-u.ac.jp
† This paper is an extended version of our paper published in the International Conference on Communications, Signal Processing and their Applications (ICCSPA '20), Sharjah United Arab Emirates, 16–18 March 2021.

Abstract: Recently, automatic speech recognition (ASR) and visual speech recognition (VSR) have been widely researched owing to development in deep learning. Most VSR research works focus only on frontal face images. However, assuming real scenes, it is obvious that a VSR system should correctly recognize spoken contents from not only frontal but also diagonal or profile faces. In this paper, we propose a novel VSR method that is applicable to faces taken at any angle. Firstly, view classification is carried out to estimate face angles. Based on the results, feature extraction is then conducted using the best combination of pre-trained feature extraction models. Next, lipreading is carried out using the features. We also developed audio-visual speech recognition (AVSR) using the VSR in addition to conventional ASR. Audio results were obtained from ASR, followed by incorporating audio and visual results in a decision fusion manner. We evaluated our methods using OuluVS2, a multi-angle audio-visual database. We then confirmed that our approach achieved the best performance among conventional VSR schemes in a phrase classification task. In addition, we found that our AVSR results are better than ASR and VSR results.

Keywords: visual speech recognition; multi-angle lipreading; automatic speech recognition; audio-visual speech recognition; deep learning; view classification

Citation: Isobe, S.; Tamura, S.; Hayamizu, S.; Gotoh, Y.; Nose, M. Multi-Angle Lipreading with Angle Classification-Based Feature Extraction and Its Application to Audio-Visual Speech Recognition. *Future Internet* **2021**, *13*, 182. https://doi.org/10.3390/fi13070182

Academic Editors: Khalid Elgazzar, Aboelmagd Noureldin, Mohamed El-Tarhuni and Mohamed Hassan

Received: 1 June 2021
Accepted: 13 July 2021
Published: 15 July 2021

1. Introduction

Recently, automatic speech recognition (ASR) has been confirmed to have high recognition performance by using deep learning (DL), an attractive artificial intelligence technology, and is used in various scenarios, such as voice input for mobile phones and car navigation systems. However, there is a problem that speech waveforms are degraded by audio noise in real environments, reducing the accuracy of speech recognition. In order to overcome this issue, we need to develop robust ASR systems against any audio noise. One of these ASR systems applicable in noisy environments is audio visual speech recognition (AVSR, also known as multi-modal speech recognition), which employs ASR frameworks with visual speech recognition (VSR, also known as lipreading). VSR uses lip images which are not affected by audio noise and estimates what a subject uttered only from a temporal sequence of lip images. VSR and AVSR have a potential to be applied in various practical applications such as automatic conference minute generation and human interfaces on smartphones. Owing to state-of-the-art DL technology, recently, we have achieved high performance of VSR. However, VSR still has several problems when we employ the technique in real-world scenes; for example, most VSR studies have only considered frontal faces, but VSR technology for non-frontal views is also essential for real applications. In other words,

assuming real scenes, a speaker does not always face a camera, such as smart device or tablet device, in a VSR or an AVSR system. We thus have been developing multi-angle VSR architecture which enables us to perform VSR when not only frontal lip images but also non-frontal lip images are observed.

There are two main approaches for multi-angle VSR. The first method is to build a VSR model using training lip images captured at several angles. The second approach is to convert non-frontal lip images to frontal ones and apply the conventional frontal VSR technique. In this paper, we focus on the first approach, and propose a feature integration-based multi-angle VSR system using DL, particularly 3D convolutional neural networks (CNNs), that are one kind of deep neural networks (DNNs). Based on most conventional multi-angle VSR studies, it is necessary to estimate at which angle lip images are captured, to choose a suitable angle-specific VSR model. However, if the system fails to estimate the right angle, the recognition performance drastically decreases. We need to build a VSR technique that can be applied to real scenes where it is difficult to estimate the accurate lip angle.

Therefore, we employ a new multi-angle VSR method, in which all angle-specific VSR models are trained using images at different angles. Our multi-angle VSR method consists of three parts: a view classification part, a feature extraction part and a recognition part. Assume that we have a sequence of lip images to be recognized. Firstly, in the view classification part, we prepare a common 2D CNN that estimates the angle of the input image (see Section 3.1.1). The model is then applied to each image in the sequence, followed by determining the angle which has the majority in the estimation. Secondly, in the feature extraction part, we build 3D CNN models for possible combinations of angle-specific training data sets (see Section 3.1.2). Based on the angle obtained in the first part, we choose the best models and extract features from the models. In the last integration part, we concatenate these features, followed by recognition by means of a fully connected (FC) neural network (see Section 3.1.3) In addition, we perform a decision fusion-based AVSR employing our proposed multi-angle VSR.

We conducted evaluation experiments using the open data set OuluVS2, in which subjects were captured simultaneously at five angles in addition to speech data. The experimental results show that our proposed method can improve VSR accuracy much more than conventional schemes on average, and achieve significant AVSR accuracy in noisy environments. In addition, we confirm that our proposed method is sufficiently robust against view classification errors, because, in the second part, we simultaneously employ several models built using multi-angle training data.

The rest of this paper is organized as follows. In Section 2, we briefly review related works on multi-angle VSR. Section 3 introduces our method. The experimental setup, results and discussion are described in Section 4. Finally, Section 5 concludes this paper.

2. Related Work

Recently, many researchers have proposed deep learning-based AVSR and VSR schemes [1–21]. As mentioned, most conventional VSR research has focused on frontal face images, assuming that VSR systems are in front of speakers, since there are only a few data sets available with multi-angle faces. Here, we introduce several lipreading works focusing not only on frontal but also diagonal and profile images. To develop these schemes, we need a research corpus. One of the public multi-angle VSR data sets is OuluVS2 [22].

An early work of multi-angle lipreading is [1], where a system was trained using either frontal (0°) or profile (90°) faces. According to the experimental results, the frontal view showed a lower word error rate (WER) than the profile view. In [2], the authors built a multi-angle system investigating a frontal (0°) view, a left profile (90°) view and a right profile (−90°) view. They reported significantly better performance when using the frontal view than the others. Saitoh et al. proposed a novel sequence image representation method called concatenated frame image (CFI) [3]. Two types of data augmentation methods for CFI, and a framework of a CFI-based CNN, were tested. Bauman et al. indicated that

human lipreaders tend to have higher performance when slightly angled faces are available, presumably because of the visibility of lip protrusion and rounding [4]. In [5], the active appearance model (AAM) was utilized for feature extraction at five angles, and lipreading was examined on a view-dependent system, as well as on a view-independent system using a regression method in a feature space. As a result, the view-dependent system performed the best performance at 30° in all tests. Zimmermann et al. used principal component analysis (PCA)-based convolutional networks together with Long short-term memories (LSTMs), one of the DL models, in addition to a conventional speech recognition model, hidden Markov models (HMMs) with Gaussian mixture models (GMMs) [6]. They aimed at combining multiple views by employing these techniques. They finally confirmed that the highest performance was obtained at 30°. Anina et al. stated that the highest accuracy was achieved at 60° in their experiments [22]. Kumar et al. showed that profile-view lipreading provides significantly lower WERs than frontal-view lipreading [7].

There is another strategy to conduct transformation to images or incorporate several views with DL technology. There is one work [8] that involved converting faces viewed from various directions to frontal faces using AAMs. The experimental results showed that recognition accuracy was improved even when the face direction changed about 30° relative to a frontal view. In [9], the authors proposed a scheme called "View2View" using an encoder–decoder model based on CNNs. The method transformed non-frontal mouth region images into frontal ones. Their results showed that the view-mapping system worked well for VSR and AVSR. Estellers et al. introduced a pose normalization technique and performed speech recognition from multiple views by generating virtual frontal views from non-frontal images [10]. In [11], Petridis et al. proposed an end-to-end multi-view lipreading system based on bidirectional LSTM networks. This model simultaneously extracted features directly from the pixels and performed visual speech classification from multi-angle views. The experimental results demonstrated that the combination of frontal and profile views improved accuracy over the frontal view. Zimmermann et al. also proposed another decision fusion-based lipreading model [12]; they extracted features through a PCA-based convolutional neural network, LSTM network and GMM–HMM scheme. The decision fusion succeeded by combining Viterbi paths. In [13], Sahrawat et al. extended a hybrid attention-based connectionist temporal classification system with view-temporal attention to perform multi-angle lipreading. Lee et al. trained an end-to-end CNN–LSTM model [14].

Many studies have been conducted focusing on AVSR. In this paper, we would like to introduce a couple of state-of-the-art works. An AVSR system based on a recurrent neural network transducer architecture was built in [15]. The authors evaluated the system using the LRS3-TED data set, achieving high performance. In [16], the authors proposed a multimodal attention-based method for AVSR, which could automatically learn fused representations from both modalities based on their importance. They employed sequence-to-sequence architectures, and confirmed high recognition performance under both acoustically clean and noisy conditions. Another AVSR system using a transformer-based architecture was proposed in [17]. The experimental results show that on the How2 data set, the system improved word error rate relatively over sub-word prediction models. In [18], we proposed an AVSR method based on deep canonical correlation analysis (DCCA). DCCA consequently generates projections from two modalities into one common space, so that the correlation of projected vectors could be maximized. We thus employed DCCA techniques with audio and visual modalities to enhance the robustness of ASR. As a result, we confirmed that DCCA features of each modality can be improved compared to the original features, and better ASR results in various noisy environments can be obtained.

Although we can find a lot of VSR and AVSR methods, there are only a few works combining ASR and multi-angle VSR to accomplish angle-invariant AVSR. One of them is [19], where the authors proposed an early fusion-based AVSR method using bidirectional LSTMs. Similar to their past work [11], the authors put lip images at various angles and corresponding audio signals into the bidirectional LSTM models.

3. Methodology

Our proposed multi-angle VSR method consists of three parts: a view classification part, a feature extraction part and a recognition part. Figure 1 depicts the architecture of our AVSR approach, including ASR and the VSR model. In this section, we describe each part of our multi-angle VSR scheme followed by ASR and AVSR frameworks.

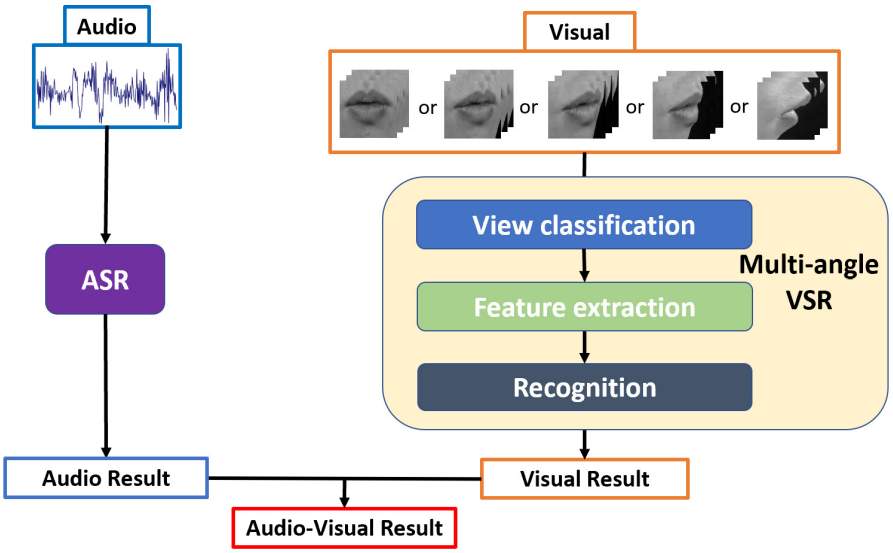


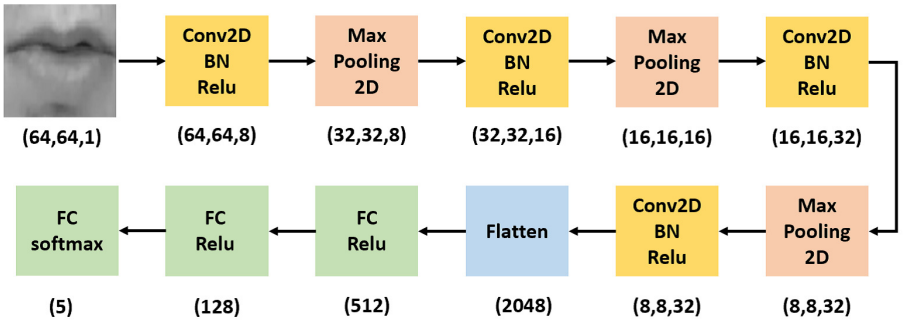
Figure 1. An architecture of our proposed multi-angle AVSR method.

3.1. Multi-Angle VSR

VSR accepts a temporal sequence of lip images to recognize what a subject utters according to the given images. Assuming real scenes, it is not guaranteed that a speaker is strictly facing a VSR system. One way to deal with this problem is to prepare several models, each of which corresponds to a certain angle, estimate at which angle face images are captured and apply a corresponding angle-specific model.

3.1.1. View Classification

In the view classification part, we at first estimate at which angle face images were recorded among the following five candidates in this work: $0°, 30°, 45°, 60°$ and $90°$. The estimation was carried out for each lip image in one sequence, using the 2D CNN model illustrated in Figure 2. The 2D CNN model employs a simple and common architecture; convolutional and pooling layers are repeatedly applied followed by FC layers, to obtain a classification result. After processing the above step for all the input images, we determine the angle which is the most often chosen.



* BN = Batch Normalization, FC = Fully Connected Layer

Figure 2. A 2D CNN model for view classification. Numbers in parentheses mean data shapes. For example, (64,64,1) indicates 64 × 64 (image size) × 1 (channel).

3.1.2. Feature Extraction

Before conducting feature extraction, we prepare 3D CNN pre-recognition models for all possible combinations of the above five angles, i.e., models each trained only using images obtained from a single angle, such as a model from frontal images and a model from 30° images, as well as models each built using data of several angles, such as a model trained using both 0° and 30° data and a model using all face images. An architecture of our 3D CNN-based VSR models is shown in Figure 3. The last layer has 20 outputs, each of which corresponds to one class in our recognition task. As a result, we build 31 models in this case ($\sum_{i=1}^{5} {}_5C_i = 5 + 10 + 10 + 5 + 1 = 31$), as shown in Table 1. Table 1 also indicates preliminary VSR results: recognition accuracy to validation data at a certain angle, using a certain model chosen among those 31 models. For example, if we adopt a 30° model for 60° data, the accuracy is 87.55%.

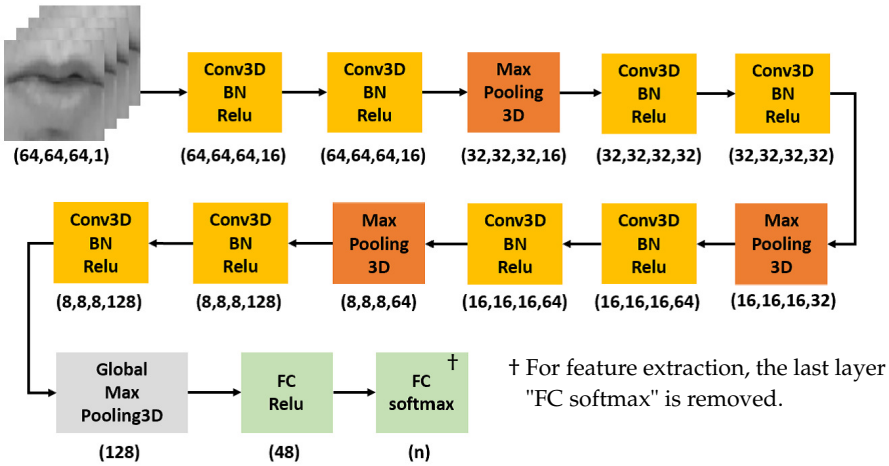


Figure 3. A 3D CNN pre-recognition model for feature extraction. Numbers in parentheses mean data shapes. For example, (64,64,64,1) indicates 64 (images) × 64 × 64 (size) × 1 (channel).

According to the angle obtained in the view classification part, we select the most reliable three models for the estimated angle, which are shown in bold in Table 1. For instance, we adopt (1) "0° + 30° + 45°", (2) "0° + 30° + 45° + 60°" and (3) "0° + 30° + 45° + 90°" models for 45° data. In other words, we determine suitable angle combination patterns of training data for the estimated angle. We then utilize those models as feature extractors; we remove the last layer, resulting in a new output layer generating a 48-dimensional feature vector, as indicated in Figure 3. Finally, we obtain three 48-dimensional vectors from this part.

This strategy has two advantages. First, as shown in Table 1, models trained using data of several angles have relatively higher performance than those trained using single angle data. This result motivates us to choose such models for multi-angle data. Second, even if the view classification fails, it is still expected to obtain high performance by our scheme; for instance, in the case where a 30° sequence is misclassified as 45°, the above models (1)∼(3) are used for feature extraction, all in which 30° data are also used in model training. There is another reason to encourage us to choose this framework. The model trained using all data, indicated in the bottom row in Table 1, achieved good performance. On the other hand, there exists a better model in all the angle cases. This suggests using only the model with all data is not the best solution. Hence, for each angle, we prepare several models trained using multi-angle data and utilize them as feature extractors.

3.1.3. Recognition

In the integration part, firstly, we integrate those 48-dimensional features extracted from three angle-specific models, by simply concatenating them. Thereafter, we conduct recognition using two FC layers ($48 \times 3 \rightarrow 48 \rightarrow 20$). Here, we apply a 50% dropout between the FC layers.

Table 1. Preliminary visual recognition accuracy (%) for validation data.

Model \ Data	0°	30°	45°	60°	90°
0°	95.33	93.33	89.78	69.22	42.78
30°	93.78	95.89	94.67	87.55	69.00
45°	88.22	91.89	95.00	93.78	76.89
60°	66.00	80.11	88.22	95.89	90.89
90°	47.44	56.55	69.44	93.56	94.67
0° + 30°	96.00	96.56	96.22	88.67	66.56
0° + 45°	94.78	95.78	95.78	93.67	79.34
0° + 60°	92.78	94.00	93.55	95.44	88.22
0° + 90°	96.33	96.67	94.67	96.56	93.56
30° + 45°	93.56	95.56	95.22	90.89	79.00
30° + 60°	93.78	96.89	96.44	97.11	87.33
30° + 90°	94.67	97.22	96.78	95.78	**95.11**
45° + 60°	88.33	92.22	96.00	96.11	89.56
45° + 90°	89.11	93.67	94.67	96.78	94.67
60° + 90°	75.11	83.56	88.00	96.89	94.45
0° + 30° + 45°	96.89	97.55	**97.89**	96.78	76.44
0° + 30° + 60°	96.11	97.78	96.89	96.67	87.56
0° + 30° + 90°	95.11	97.89	96.78	95.89	94.45
0° + 45° + 60°	96.11	96.89	96.00	97.22	85.67
0° + 45° + 90°	94.33	95.78	95.44	95.44	93.56
0° + 60° + 90°	96.22	96.78	95.56	97.55	94.78
30° + 45° + 60°	93.78	96.89	97.44	96.55	84.11
30° + 45° + 90°	95.78	97.11	97.22	97.22	94.78
30° + 60° + 90°	95.67	97.78	97.33	97.56	94.56
45° + 60° + 90°	89.89	92.89	95.33	96.55	94.66
0° + 30° + 45° + 60°	96.67	96.89	**97.78**	97.11	86.33
0° + 30° + 45° + 90°	**97.44**	**98.33**	**97.89**	**97.67**	95.00
0° + 30° + 60° + 90°	96.44	**98.11**	97.00	**98.11**	94.22
0° + 45° + 60° + 90°	**97.67**	**98.22**	97.45	**98.22**	93.89
30° + 45° + 60° + 90°	95.22	96.78	97.11	97.22	**96.45**
0° + 30° + 45° + 60° + 90°	**96.89**	97.89	97.00	97.55	**95.89**

3.2. ASR

3.2.1. Feature Extraction

In our ASR framework, we extract 13 mel-frequency cepstrum coefficients (MFCCs) in addition to 13 ΔMFCCs and 13 ΔΔMFCCs from audio waveforms with a frame length of 25 msec and a frame shift of 10 msec [23–26]. The MFCC is the most commonly used feature in the speech recognition field in addition to ΔMFCCs and ΔΔMFCCs, which are first and second derivatives, respectively. As a result, we obtain a 39-dimensional acoustic vector.

In the acoustic modality, there are many frameworks and a lot of features, e.g., [27, 28]. We should carefully choose an audio processing scheme based on performance and theoretical perspectives. For instance, mel-frequency spectrograms are commonly used for CNN-based speech recognition. In this study, we first conduct preliminary experiments to measure the accuracy when using mel-frequency spectrograms or MFCCs. The size of the spectrograms is 96 × 128. Because using MFCCs with CNNs achieves better performance, we choose this framework. Note that we need to investigate which acoustic processing methods and features are the most suitable for the other tasks.

3.2.2. Recognition

After computing MFCCs from consecutive frames, we apply a 2D CNN-based model for recognition, which is illustrated in Figure 4. Similar to the VSR model, we finally obtain an audio result including a probability for each class.

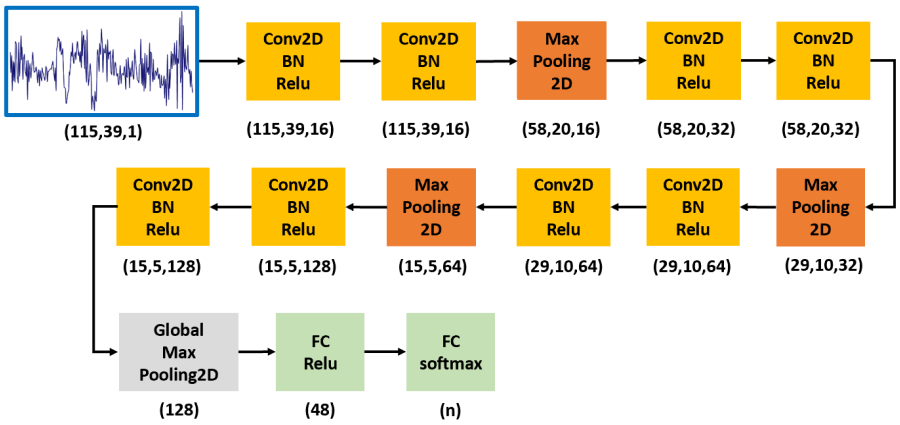


Figure 4. A 2D CNN model for ASR. Numbers in parentheses mean data shapes. For example, (115,39,1) indicates 115 (vectors) × 39 (MFCCs) × 1 (channel).

3.3. AVSR

Firstly, a sequence of lip images is added to the VSR model, while corresponding speech data are given to the ASR model. As mentioned in detail later, we adopt the corpus OuluVS2, in which the task is to estimate which sentence is spoken. Therefore, for each class, we obtain a probability from ASR results and another one from VSR. These probabilities are integrated in a decision fusion manner. Let us denote conditional probabilities of class c from ASR and VSR models by $P_A(\mathbf{x}_A|c)$ and $P_V(\mathbf{x}_V|c)$, respectively. Here, $\mathbf{x}_A$ indicates an audio input representation, and $\mathbf{x}_V$ means the corresponding image vector. We then obtain an audio-visual probability $P_{AV}(\mathbf{x}_A, \mathbf{x}_V|c)$ as:

$$P_{AV}(\mathbf{x}_A, \mathbf{x}_V|c) = \alpha P_A(\mathbf{x}_A|c) + (1-\alpha)P_V(\mathbf{x}_V|c) \quad (1)$$

In this work, we simply fix $\alpha = 0.5$.

4. Experiments

In order to examine the effectiveness of our VSR scheme as well as AVSR framework, we carry out recognition experiments.

4.1. Data Set

4.1.1. OuluVS2

We choose the OuluVS2 corpus to evaluate our scheme. The database contains 10 short phrases, 10 digits sequences and 10 TIMIT sentences uttered by 52 speakers. The corpus includes face images captured by five cameras simultaneously at 0° (frontal), 30°, 45°, 60° and 90° (profile) angles. In this study, we adopt the phrase data and digit data, uttered three times by each speaker. In our experiment, the data spoken by 52 speakers are divided into training data by 35 speakers (speaker ID:1–36), validation data by 5 speakers (speaker ID: 37–41) and testing data by 12 speakers (speaker ID: 42–53). Note that the speaker ID: 29 is missing. We conduct the same data split as previous works, such as [3,6,14], for a fair comparison. We also check whether the data split is appropriate by changing the different split settings, and confirm that using the data sets gives us fair results. The phrases are as follows: "Excuse me", "Goodbye", "Hello", "How are you", "Nice to meet you", "See you", "I am sorry", "Thank you", "Have a good time", "You are welcome". Each digit utterance consists of 10 digits randomly chosen. Note that, since we use a part of this

corpus to enhance model training data, the task in this work is a 10-class classification for phrase utterances.

4.1.2. DEMAND

We select another database, DEMAND [29], as a noise corpus. This corpus consists of six primary categories, each of which has three environments. Four of those primary categories are for closed spaces: Domestic, Office, Public and Transportation. The remaining two categories are recorded outdoors: Nature and Street. In this study, we add some of those noises to build audio training data.

4.1.3. CENSREC-1-AV

CENSREC-1-AV [30] is a Japanese audio-visual corpus for noisy multi-modal speech recognition. CENSREC-1-AV provides audio utterances, lip images and audio noise. In this study, we utilize the audio noise, i.e., interior car noises recorded on city roads and expressways, to obtain acoustically noisy testing data.

4.2. Experimental Setup

We evaluate a model by utterance-level accuracy:

$$\text{Accuracy} = \frac{H}{N} \times 100\ [\%] \tag{2}$$

where H and N are the number of correctly recognized utterances and the total number of utterances, respectively. In addition, we also evaluate our model performance by the F1 score. An F1 score can be computed as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

where

$$\text{Precision} = \frac{T_P}{T_P + F_P} \quad , \quad \text{Recall} = \frac{T_P}{T_P + F_N} \tag{4}$$

In Equation (4), T_P is the number of correctly classified utterances. F_P and F_N indicate false positives and false negatives, respectively. We calculate the score in each class.

Since DNN-based model performance slightly varies depending on the probabilistic gradient descent algorithm, which is a common model training approach, we repeat the same experiment three times and the mean accuracy is calculated. In terms of DNN hyperparameters, we choose a cross-entropy function as a loss function and Adam as an optimizer. Batch size, epochs and learning rate are set to 32, 50 and 0.001, respectively. We carry out our experiments using NVIDIA GEFORCE RTX 2080 Ti.

4.3. Preprocessing

The OuluVS2 data set includes extracted lip images, however, the image size is not consistent. In order to apply DNNs, we resize all images to 64 $\times$ 64. Based on our preliminary experiments with different image sizes, considering classification accuracy and computational cost, we use the image size of 64 $\times$ 64. Furthermore, we normalize a frame length to 64; if the length is less than 64 we conduct upsampling, otherwise we suppress some frames. In addition, we convert all color images to gray-scale ones. Similar to visual frames, we normalize the audio frame length to 115; if the length is less than 115 we add last frame, otherwise up to 115 frames are used.

In the OuluVS2 corpus, there are 1050 (35 speakers $\times$ 10 utterance $\times$ 3 times) sentences available. However, the data size is not enough for DNN model training. To compensate for the lack of training data, we apply data augmentation in the audio and visual modalities. In the audio modality, we add acoustic noises in DEMAND to the original utterance data. The details, including noise type and signal-to-noise Ratio (SNR) conditions, are shown in

Table 2. In the visual modality, we train our VSR models using not only phrase data but also digit sequence data based on our previous work [31].

Table 2. The amount of acoustically noisy data for ASR training: e.g., 1050 = 35 (speakers) × 10 (utterances) × 3 (times).

Noise \ SNR	0 dB	5 dB	10 dB	15 dB	20 dB
Kitchen	1050	-	1050	-	1050
Park	-	1050	-	1050	-
Office	1050	-	1050	-	1050
Station	-	1050	-	1050	-
Car	1050	-	1050	-	1050

4.4. Results and Discussion

4.4.1. View Classification

First of all, we investigated view classification performance. View classification results for the test data are shown in Table 3. The whole accuracy of view classification was 91.39%. Focusing on the results for each angle, classification for frontal and profile views was fully successful. On the other hand, misclassification was found in the diagonal views, particularly at 45°. In conclusion, the performance of our view classification was acceptable. However, the last fact also indicates that it is required for the following VSR models to carry out recognition successfully even for the miscategorized sequences.

Table 3. A confusion matrix of view classification results.

Result \ Label	0°	30°	45°	60°	90°
0°	360	0	0	0	0
30°	0	324	18	0	0
45°	0	36	242	0	0
60°	0	0	100	359	0
90°	0	0	0	1	360

4.4.2. VSR

Recognition accuracy of our and competitive VSR schemes is shown in Table 4. We firstly tested our models with and without view classification. Our method with the view classification part achieved almost the same or better performance, compared to ours without the classification, in which the classification result was correctly given. This indicates our feature extraction and recognition strategy can perform well.

Table 4. VSR accuracy (%) of our proposed method and conventional schemes.

Method \ Data	0°	30°	45°	60°	90°	Mean
CNN + Data Augmentation [3]	85.6	82.5	82.5	83.3	80.3	82.84
PCA + LSTM + GMM–HMM [6]	73.1	75.6	67.2	63.3	59.3	67.7
View2View [9]	-	86.11	83.33	81.94	78.89	82.57
End-to-end Encoder + BLSTM [11]	**91.8**	87.3	88.8	86.4	**91.2**	89.1
End-to-End CNN–LSTM [14]	82.8	81.1	85.0	83.6	86.4	83.78
Ours without view classification	91.02	90.56	91.20	90.00	88.88	90.33
Ours with view classification	91.02	**91.38**	**92.21**	**90.09**	88.88	**90.65**

Next, we compared our approach with conventional methods. Focusing on the average of recognition accuracy, our proposed method achieved the highest accuracy regardless of the presence or absence of the view classification part. It is interesting that at 45° we found much more improvement than in the other conditions, and even the view classification performance was insufficient. Since 45° data were used as training data in the neighboring 30° and 60° conditions, we might obtain such an improvement even if the view classification fails. We also found that our method was particularly effective in the medium-angle (30°, 45° and 60°) conditions, while the end-to-end system had higher accuracy for frontal and profile images.

Figure 5 indicates F1 scores for each angle. Among all the angles, it is found that shorter utterances were relatively hard to classify, because there were fewer cues for recognition.

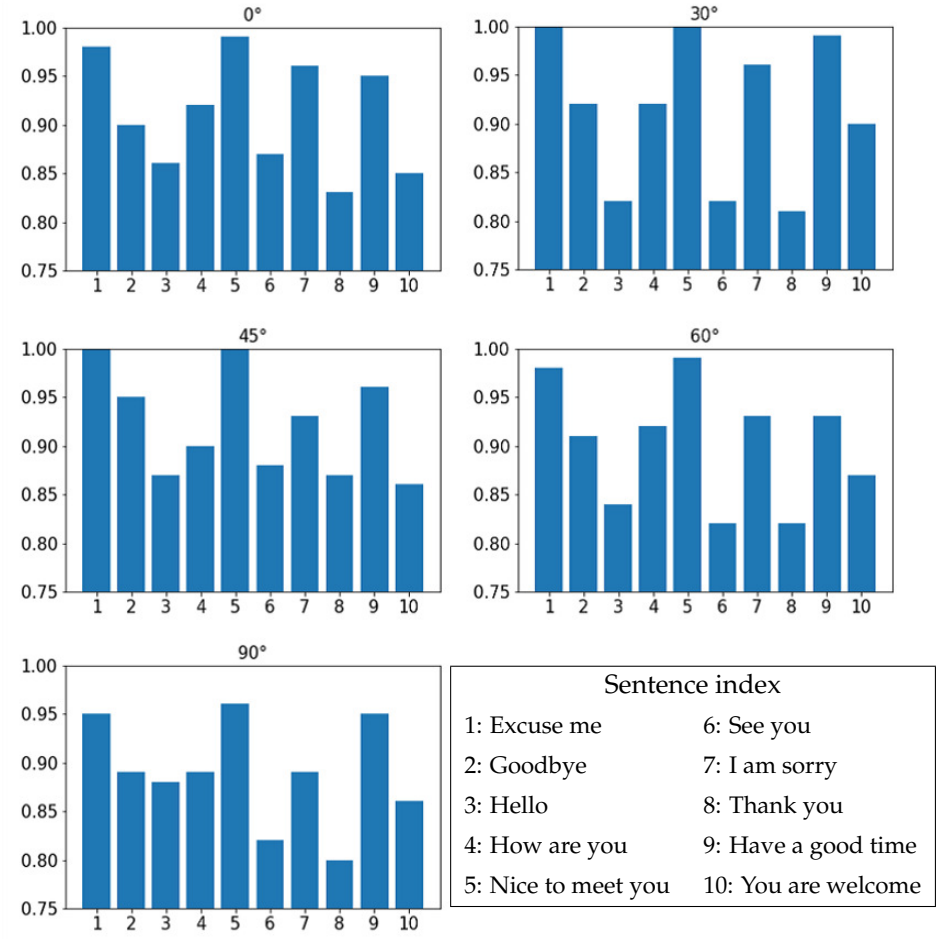


Figure 5. F1 scores in every class of our proposed multi-angle VSR model. The vertical axis indicates the F1 score, and the horizontal number means a sentence index.

4.4.3. AVSR

Table 5 shows recognition accuracy of our ASR, VSR and AVSR methods in various noise environments. Note that, because the task was a 10-class classification, the accuracy in noisy environments tended to be higher compared to large-vocabulary speech recognition. The VSR accuracy was stable and unrelated to SNR since visual information is not affected by noise. As is already known, the results of VSR were lower than those of ASR in all the SNRs, because audio features are more effective and informative than visual ones. Among the models, AVSR achieved the best accuracy in all the conditions. In particular, at 0 dB, where the effect of noise was the largest, the performance was improved by 3% for city road noise and by 2.3% for expressway noise compared to ASR results. Even in the case of

20 dB, where the effect of noise was quite small, the accuracy was slightly improved. As mentioned, we employed the decision fusion strategy, which is the simplest integration method. Similar to the ensemble approach, we believe our decision fusion method could successfully integrate ASR and VSR results, which had different recognition errors.

Table 5. AVSR accuracy (%) in various noise conditions with ASR and VSR.

Model \ Data	Noise	0 dB	5 dB	10 dB	15 dB	20 dB
ASR	city road	95.83	99.26	99.26	99.35	99.26
VSR				90.65		
AVSR		**98.70**	**99.63**	**99.72**	**99.63**	**99.63**
ASR	expressway	96.85	99.44	99.35	99.26	99.35
VSR				90.65		
AVSR		**99.17**	**99.72**	**99.53**	**99.72**	**99.72**

5. Conclusions

In this paper, we proposed a multi-angle VSR system in which feature extraction was conducted using angle-specific models based on view classification results, followed by feature integration and VSR. We also proposed a decision fusion-based AVSR. We employed DNNs in our system, to perform view classification, feature extraction and recognition. The advantages of our method are choosing appropriate feature extraction models based on angle classification results, reducing the negative impact of misclassification, and incorporating ASR and VSR results efficiently. Evaluation experiments were conducted using the multi-view corpus OuluVS2. Then, we found our scheme could work well compared to past works, and we clarified the effectiveness of view classification and feature extraction from pre-trained angle-specific models. Moreover, we found that our AVSR method is superior to ASR and VSR because our decision fusion method could successfully integrate ASR and VSR results.

As our future work, we are planning to conduct experiments using different angle settings and other tasks. The implementation of this framework for real applications is also expected. In addition, because there are some research works investigating spectrograms instead of MFCCs, we will try to employ spectrograms as acoustic input. Finally, we will explore the suitable model architecture and its physical meaning for feature extraction.

Author Contributions: Funding acquisition, Y.G. and M.N.; Investigation, S.I.; Methodology, S.I.; Project administration, S.T.; Supervision, S.H.; Writing – original draft, S.I.; Writing – review & editing, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: The databases used in this article are OuluVS2, DEMAND and CENSREC-1-AV. For details, please refer to [22], [29] and [30], respectively.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lucey, P.; Potamianos, G. Lipreading using profile versus frontal views. In Proceedings of the MMSP, Victoria, BC, Canada, 3–6 October 2006; pp. 24–28.
2. Lucey, P.; Sridharan, S.; Dean, D. Continuous pose invariant lipreading. In Proceedings of the INTERSPEECH, Brisbane, Australia, 22–26 September 2008; pp. 2679–2682.
3. Saitoh, T.; Zhou, Z.; Zhao, G.; Pietikäinen, M. Concatenated frame image based CNN for visual speech recognition. In Proceedings of the ACCV, Taipei, Taiwan, 21–23 November 2016.
4. Bauman, S.L.; Hambrecht, G. Analysis of view angle used in speech reading training of sentences. *Am. J. Audiol. 1995*, *4*, 67–70.
5. Lan, Y.; Theobald, B.J.; Harvey, R. View independent computer lip-reading. In Proceedings of the Multimedia and Expo, Melbourne, Australia, 9–13 July 2012; pp. 432–437.

6. Zimmermann, M.; Ghazi, M.M.; Ekenel, H.K.; Thiran, J.-P. Visual speech recognition Using PCA networks and LSTMs in a tandem GMM-HMM system. In Proceedings of the ACCV, Taipei, Taiwan, 21–23 November 2016.
7. Kumar, K.; Chen, T.; Stern, R. Profile view lip reading. In Proceedings of the ICASSP, Honolulu, HI, USA, 15–20 April 2007; pp. 429–432.
8. Komai, Y.; Yang, N.; Takiguchi, T.; Ariki, Y. Robust AAM based audio-visual speech recognition against face direction changes. In Proceedings of the ACM Multimedia, Nara, Japan, 29 October–2 November 2012; pp. 1161–1164.
9. Koumparoulis, A.; Potamianos, G. Deep view2view mapping for view-invariant lipreading. In Proceedings of the SLT, Athens, Greece, 18–21 December 2018; pp. 588–594.
10. Estellers, V.; Thiran, J.-P. Multipose audio-visual speech recognition. In Proceedings of the EUSIPCO, Barcelona, Spain, 29 August–2 September 2011; pp. 1065–1069.
11. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-end multiview lip reading. In Proceedings of the ICASSP, Calgary, AB, Canada, 15–20 April 2018; pp. 6548–6552.
12. Zimmermann, M.; MehdipourGhazi, M.; Ekenel, H.K.; Thiran, J.-P. Combining multiple views for visual speech recognition. In Proceedings of the AVSP, Stockholm, Sweden, 25–26 August 2017.
13. Sahrawat, D.; Kumar, Y.; Aggarwal, S.; Yin, Y.; Shah, R.R.; Zimmermann, R.; "Notic My Speech"—Blending Speech Patterns With Multimedia. *arXiv* **2020**, arXiv:2006.08599.
14. Lee, D.; Lee, J.; Kim, K.E. Multi-view automatic lip-reading using neural network. In Proceedings of the ACCV, Taipei, Taiwan, 21–23 November 2016.
15. Makino, T.; Liao, H.; Assael, Y.; Shillingford, B.; Garcia, B.; Braga, O.; Siohan, O. Recurrent Neural Network Transducer for Audio-Visual Speech Recognition. *arXiv* **2019**, arXiv:1911.04890v1.
16. Zhou, P.; Yang, W.; Chen, W.; Wang, Y.; Jia, J. Modality Attention for End-to-End Audio-visual Speech Recognition. *arXiv* **2019**, arXiv:1811.05250v2.
17. Paraskevopoulos, G.; Parthasarathy, S.; Khare, A.; Sundaram, S. Multiresolution and Multimodal Speech Recognition with Transformers. *arXiv* **2020**, arXiv:2004.14840v1.
18. Isobe, S.; Tamura, S.; Hayamizu, S. Speech Recognition using Deep Canonical Correlation Analysis in Noisy Environments. In Proceedings of the ICPRAM, Online, 4–6 February 2021; pp. 63–70.
19. Petridis, S.; Wang, Y.; Li, Z.; Pantic, M. End-to-End Audiovisual Fusion with LSTMs. In Proceedings of the ICASSP, Calgary, AB, Canada, 15–20 April 2018.
20. Lee, Y.H.; Jang, D.W.; Kim, J.B.; Park, R.H.; Park, H.M. Audio–visual speech recognition based on dual cross-modality attentions with the transformer model. *Appl. Sci.* **2020**, *10*, 7263.
21. Bear, H.L.; Harvey, R. Alternative visual units for an optimized phoneme-based lipreading system. *Appl. Sci.* **2019**, *9*, 2019.
22. Anina, I.; Zhou, Z.; Zhao, G.; Pietikäinen, M. OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis. In Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015.
23. Tuasikal, D.A.A.; Nugraha, M.B.; Yudhatama, E.; Muharom, A.S.; Pura, M. Word Recognition for Color Classification Using Convolutional Neural Network. In Proceedings of the CONMEDIA, Bali, Indonesia, 9–11 October 2019.
24. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-End Audiovisual Speech Recognition. In Proceedings of the ICASSP, Calgary, AB, Canada, 15–20 April 2018.
25. Mahmood, A.; Köse, U. Speech recognition based on convolutional neural networks and MFCC algorithm. In Proceedings of the AAIR, Uttar Pradesh, India, 30 June–4 July 2021; Volume 1, pp. 6–12.
26. Kathania, H.K.; Shahnawazuddin, S.; Adiga, N.; Ahmad, W. Role of Prosodic Features on Children's Speech Recognition. In Proceedings of the ICASSP, Calgary, AB, Canada, 15–20 April 2018.
27. Bountourakis, V.; Vrysis, L.; Konstantoudakis, K.; Vryzas, N. An Enhanced Temporal Feature Integration Method for Environmental Sound Recognition. *Acoustics* **2019**, *1*, 410-422.
28. Vrysis, L.; Hadjileontiadis, L.; Thoidis, I.; Dimoulas, C.; Papanikolaou, G. Enhanced Temporal Feature Integration in Audio Semantics via Alpha-Stable Modeling. *J. Audio Eng. Soc.* **2021**, *69*, 227–237.
29. Thiemann, J.; Ito, N.; Vincent, E. DEMAND: A collection of multichannel recordings of acoustic noise in diverse environments. In Proceedings of the ICA, Montreal, QC, Canada, 2–7 June 2013. Available online: https://zenodo.org/record/1227121#.YNS2p3X7Q5k (accessed on 14 July 2021).
30. Tamura, S.; Miyajima, C.; Kitaoka, N.; Yamada, T.; Tsuge, S.; Takiguchi, T.; Yamamoto, K.; Nishiura, T.; Nakayama, M.; Denda, Y.; et al. CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition. In Proceedings of the AVSP, Kanagawa, Japan, 30 September–3 October 2010; pp. 85–88. Available online: http://research.nii.ac.jp/src/en/CENSREC-1-AV.html (accessed on 14 July 2021).
31. Isobe, S.; Tamura, S.; Hayamizu, S.; Gotoh, Y.; Nose, M. Multi-angle lipreading using angle classification and angle-specific feature integration. In Proceedings of the ICCSPA, Sharjah, United Arab Emirates, 16–18 March 2020.