Unsupervised domain adaptation for lip reading based on cross-modal knowledge distillation

Takashima, Yuki ; Takashima, Ryoichi ; Tsunoda, Ryota ; Aihara, Ryo ; Takiguchi, Tetsuya ; Ariki, Yasuo ; Motoyama, Nobuaki

(Citation)
EURASIP Journal on Audio, Speech, and Music Processing,2021(1):44

(Issue Date)
2021-12-11

(Resource Type)
journal article

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/90008868

Takashima *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:44
https://doi.org/10.1186/s13636-021-00232-5

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Unsupervised domain adaptation for lip reading based on cross-modal knowledge distillation

Yuki Takashima[1], Ryoichi Takashima[1*], Ryota Tsunoda[1], Ryo Aihara[2], Tetsuya Takiguchi[1], Yasuo Ariki[1] and Nobuaki Motoyama[2]

Abstract

We present an unsupervised domain adaptation (UDA) method for a lip-reading model that is an image-based speech recognition model. Most of conventional UDA methods cannot be applied when the adaptation data consists of an unknown class, such as out-of-vocabulary words. In this paper, we propose a cross-modal knowledge distillation (KD)-based domain adaptation method, where we use the intermediate layer output in the audio-based speech recognition model as a teacher for the unlabeled adaptation data. Because the audio signal contains more information for recognizing speech than lip images, the knowledge of the audio-based model can be used as a powerful teacher in cases where the unlabeled adaptation data consists of audio-visual parallel data. In addition, because the proposed intermediate-layer-based KD can express the teacher as the sub-class (sub-word)-level representation, this method allows us to use the data of unknown classes for the adaptation. Through experiments on an image-based word recognition task, we demonstrate that the proposed approach can not only improve the UDA performance but can also use the unknown-class adaptation data.

Keywords: Lip reading, Knowledge distillation, Multimodal, Unsupervised domain adaptation

1 Introduction

Lip reading is a technique of understanding utterances by visually interpreting the movements of a person's lips, face, and tongue when the spoken sounds cannot be heard. For example, for people with hearing problems, lip reading is one communication skill that can help them communicate better. McGurk et al. [1] reported that we human beings perceive a phoneme not only from the auditory information of the voice but also from visual information associated with the movement of the lips and face. Moreover, it is reported that we try to catch the movement of lips in a noisy environment and we misunderstand the utterance when the movements of the lips and the voice are not synchronized. Therefore, understanding the relationship between the voice and the movements of the lips is very important for speech perception. In the field of automatic speech recognition (ASR), visual information is used to assist the performance of speech recognition in a noisy environment [2]. In this work, lip reading has the goal of classifying words from the movements of the lips.

Recently, deep learning-based models have improved the performance of audio-visual automatic speech recognition (AV-ASR) or lip reading [3–7] where a large amount of training data is available. However, in a variety of real-life situations, there is often a mismatch between the training environment and the real environment where a user utilizes the system, and it is not easy to collect a sufficient amount of training data in a specific environment. Therefore, an effective way to adapt the model to a new environment is required. This is known as the domain adaptation (DA) problem.

The purpose of DA is to adapt a model trained on a source domain (source model) to a new target domain

*Correspondence: rtakashima@port.kobe-u.ac.jp
[1]Graduate School of System Informatics, Kobe University, Kobe, Japan
Full list of author information is available at the end of the article

by using a relatively small amount of additional training (adaptation) data. Especially, in the case when all the adaptation data are not labeled, it is called "unsupervised domain adaptation" (UDA). Various UDA approaches have been proposed [8–10]. However, most of them assume that all the adaptation data belong to classes that are defined in the source model. This means that we cannot use the real environment data for the adaptation if that data is out-of-class (e.g., out-of-vocabulary (OOV) words in speech recognition). For more practical adaptation, it is preferable if out-of-class (unknown class) data can also be used. With this in mind, in this paper, we investigate an unknown-class-driven UDA method. Although there has been research carried out to tackle a similar issue [10, 11], the UDA on the unknown-class data is an extremely challenging task because we cannot use any conventional training policies, such as maximizing the output probability of the correct class.

In this paper, we propose a UDA method based on a model for cross-modal knowledge distillation (KD) for lip reading. There are two key factors: cross-modal KD and its application to UDA. KD [12] was originally introduced as model compression, in which a small model (student model) is trained to imitate an already-trained larger model (teacher model). Based on the idea that KD can transfer the knowledge of the teacher model to the student model, this technique has been applied to various tasks [13–15]. In this paper, we investigate cross-modal KD, where the student and teacher model are a lip-reading model and an audio-based speech recognition model (ASR model), respectively. Our proposed method uses audio-visual data for training and adapting the lip-reading model. Before training the lip-reading model, we train the ASR model using audio data. In our research, we use an ASR model based on an artificial neural network. Then, we train the lip-reading model by using the output from the intermediate layer of the ASR model. Typically, the audio data has more information for recognizing speech and shows better recognition accuracy than the visual data. For this reason, the use of the output from the ASR model can be a powerful teacher.

Another important factor is the use of the data of the unknown class for UDA. The basic KD that minimizes the distance between the output probabilities (i.e., output of the final layer) of the teacher and student models cannot be applied to unknown class UDA because the output labels of the source model do not contain the target class. To solve this problem, we use the output of the intermediate layer for the KD instead of that of the final layer. This approach is advantageous because, unlike basic final-layer-based KD, our intermediate-layer-based KD can construct the sub-class (e.g., sub-word in speech recognition) representation implicitly inside the network. By using this sub-class representation as an adaptation objective, we can use the unknown class data for the adaptation.

Our approach, which utilizes an audio signal to enhance the lip-reading performance, is suitable for applications having a video camera, such as car navigation systems using in-vehicle cameras and service robots. In these applications, we can use both audio and video signals, and improving the lip-reading performance is expected to contribute to the improvement of audio-visual speech recognition performance. In the experiment, we demonstrate that our proposed method can improve the UDA performance on a word recognition task.

2 Related works

There have been many studies carried out on AV-ASR over the years, and most of them discuss how to integrate multimodal features [3, 15, 16]. We expect that to improve the performance of the lip reading can contribute to improving the performance of AV-ASR. LipNet [17] performs end-to-end sentence-level lip reading. This model consists of spatiotemporal convolutions and recurrent operations, and that is trained by a connectionist temporal classification loss [18]. MobiLipNet [19] has been proposed to achieve computationally efficient lip reading, and that uses the depthwise convolution and the pointwise convolution. There are some prior works based on a generative adversarial network (GAN) [20] for lip reading. Wand et al. [21] proposed a speaker-independent lip-reading system using domain-adversarial training that trains a model that can extract the speaker-invariant feature representation. Oliveira et al. [22] investigated a method to recognize viseme, that is the visual correspondent of a phoneme, using GAN-based mapping to alleviate a head-pose variation problem.

There have been some studies on cross-modal KD [15, 23, 24] for the purpose of transferring the knowledge of a modal having rich training data to a modal having poor training data. This technique has also been applied to the AV-ASR task [15], where the knowledge of the audio trained from a large amount of speech data is transferred to the AV-ASR model. In that study, they focused only on the case in which the audio data is corrupted by noise, and did not discuss the environmental mismatch in the image data, which is our target issue. For lip reading, a similar approach to our proposed method was used more recently as multi-granularity KD from a speech recognizer to a lip reader (LIBS) [25] where a frame-level KD corresponds to our intermediate-layer-based KD. In order to take account of the difference between the audio and video sampling rates, LIBS employs an attention mechanism. Different from LIBS, our method uses a pyramid structure to obtain the audio and video sequences of the same length. Moreover, our method is evaluated on a word-level recognition task, while LIBS

was evaluated on a sequence-level utterance recognition task.

The recent UDA approach involves finding a common representation for the two domains. Deep domain confusion [26] learns the meaningful and domain-invariant representation with an additional adaptive layer and loss function. GAN-based UDA approaches [27, 28] aim to learn the intermediate representation that cannot be used to distinguish the domain. Saito et al. [10] proposed a method to maximize the discrepancy between two classifier outputs considering the task-specific-design boundaries. Sohn et al. [29] proposed a feature-level UDA method using unlabeled video data that distills knowledge from a still image network to a video adaptation network. Afoura et al. [24] proposed a cross-modal KD method to improve the performance of lip reading using an ASR model. In our study, we investigate the use of cross-modal KD to adapt a model to the target environment.

Despite the recent progress of UDA, these conventional methods assume that all the adaptation data belong to classes that are defined in the source model, and none of the data can be used for the adaptation if that data is out-of-class. In the field of voice conversion, some approaches that do not require any context information have been proposed (e.g., [30]). Similar to these works, a context-independent (i.e., class-independent) approach for training the lip-reading model is required. In this work, we focus on the scenario where only the data of the unknown class is available during adaptation.

3 Proposed method

We aim to achieve UDA using the data of the unknown class on lip reading, which estimates the word label from an image input. In our proposed method, we use audio-visual data for training and adapting the lip-reading model, and for evaluating, we use only visual data. First, we explain the basic idea of the cross-modal KD on which our method is based. Then, we describe our proposed UDA method, which is based on the cross-modal KD using the data of the unknown class.

3.1 Cross-modal KD

Figure 1 shows an overview of the basic procedure of cross-modal KD, where the speech and the image are given from the same utterance. In our lip-reading task, the output is defined by the word. First, in advance, we train the audio model, which estimates the probability of the word from the acoustic feature using the cross entropy loss with the correct label. Given an acoustic feature x_{aud} and an image feature x_{vis}, the basic KD loss is defined as follows:

$$-\sum_{l} p_{\mathrm{aud}}(l|x_{\mathrm{aud}}) \ln p_{\mathrm{vis}}(l|x_{\mathrm{vis}}), \tag{1}$$

where $p_{\mathrm{vis}}(l|x_{\mathrm{vis}})$ and $p_{\mathrm{aud}}(l|x_{\mathrm{aud}})$ denote the probabilities of a label l estimated from the visual model based on x_{vis} and estimated from the audio model based on the input x_{aud}, respectively. Here, the acoustic feature and the image feature are extracted from the same utterance. When training the visual model, the parameters of the audio model are fixed. This loss function forces the visual model to imitate the outputs extracted from the audio model. Practically speaking, the softmax loss using the correct label (hard loss) is often used for stable training with the linear interpolation parameter λ. Li et al. [15] demonstrate that KD between the ASR model and the AV-ASR model improves the recognition performance when the speech data is corrupted by noise. Therefore, it is expected that KD between the audio model (ASR model) and the visual model (lip-reading model) also contributes to improving the performance in our task.

3.2 Cross-modal KD-based UDA for the unknown class

Before describing our method, we first want to highlight the fact that the adaptation data does not belong to any class of the source domain. Considering the domain, let

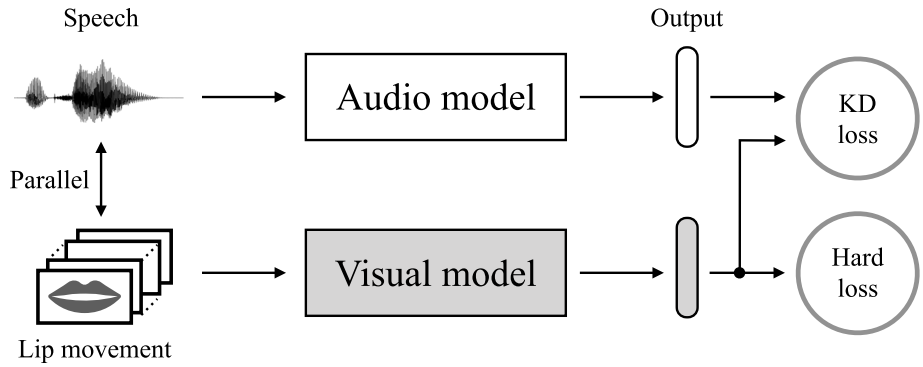

Fig. 1 Basic procedure of cross-modal knowledge distillation

$\mathcal{D}$ be the joint distribution over sequences of audio features and visual features, and the corresponding label. The output of the network is defined by the word.

Our model consists of two parts: an encoder and a classifier, as shown in Fig. 2. The encoder is a stacked convolution layer. The two encoders of the audio and visual modal can be defined as follows:

$$\boldsymbol{h}^a = f_{\text{aud}}(\boldsymbol{a}), \tag{2}$$

$$\boldsymbol{h}^v = f_{\text{vis}}(\boldsymbol{v}), \tag{3}$$

where $\boldsymbol{a} = (a_1, ..., a_t, ..., a_{T_a})$ and $\boldsymbol{v} = (v_1, ..., v_{t'}, ..., v_{T_v})$ are input sequences of acoustic features and of visual features, respectively. $\boldsymbol{h}^a = (h_1^a, ..., h_u^a, ..., h_U^a)$ and $\boldsymbol{h}^v = (h_1^v, ..., h_u^v, ..., h_U^v)$ are the sequences of high-level representations. Here a_t, $v_{t'}$, $h_u^a \in \mathbb{R}^d$, and $h_u^v \in \mathbb{R}^d$ are the input acoustic feature frame, the input visual feature frame, and the d-dimensional encoder features of both modalities, respectively. T_a, T_v and $U \leq \min(T_a, T_v)$ denote the numbers of the input acoustic features and the input visual features, and the number of the encoder output features, respectively. The number of steps of encoded features is the same between the two modalities. The classifier consists of fully connected layers to estimate the corresponding word label.

During adaptation, our method minimizes the mean square error (MSE) between the hidden representations as follows:

$$\mathcal{L}_{MSE}(\mathcal{D}) = \mathbb{E}_{\{\boldsymbol{v},\boldsymbol{a},\boldsymbol{y}\}\sim\mathcal{D}}[\,||\boldsymbol{h}^a - \boldsymbol{h}^v||_2^2\,], \tag{4}$$

where $\boldsymbol{y}$ is the label and is ignored. Unlike the generally used KD loss (Eq. (1)), we use the hidden representation in the intermediate layer for distillation. In the output layer and layers in the classifier, the frame-level information is lost and the feature representation is specialized to word-level (i.e., class-level) information. For this reason, the simple KD formulation cannot be applied to the adaptation if the adaptation data is out-of-class. On the other hand, the layers in the encoder have sub-word or phoneme-like representation that is independent of the class because they still retain the frame-level information. For this reason, our proposed method realizes UDA using the data of the unknown class.

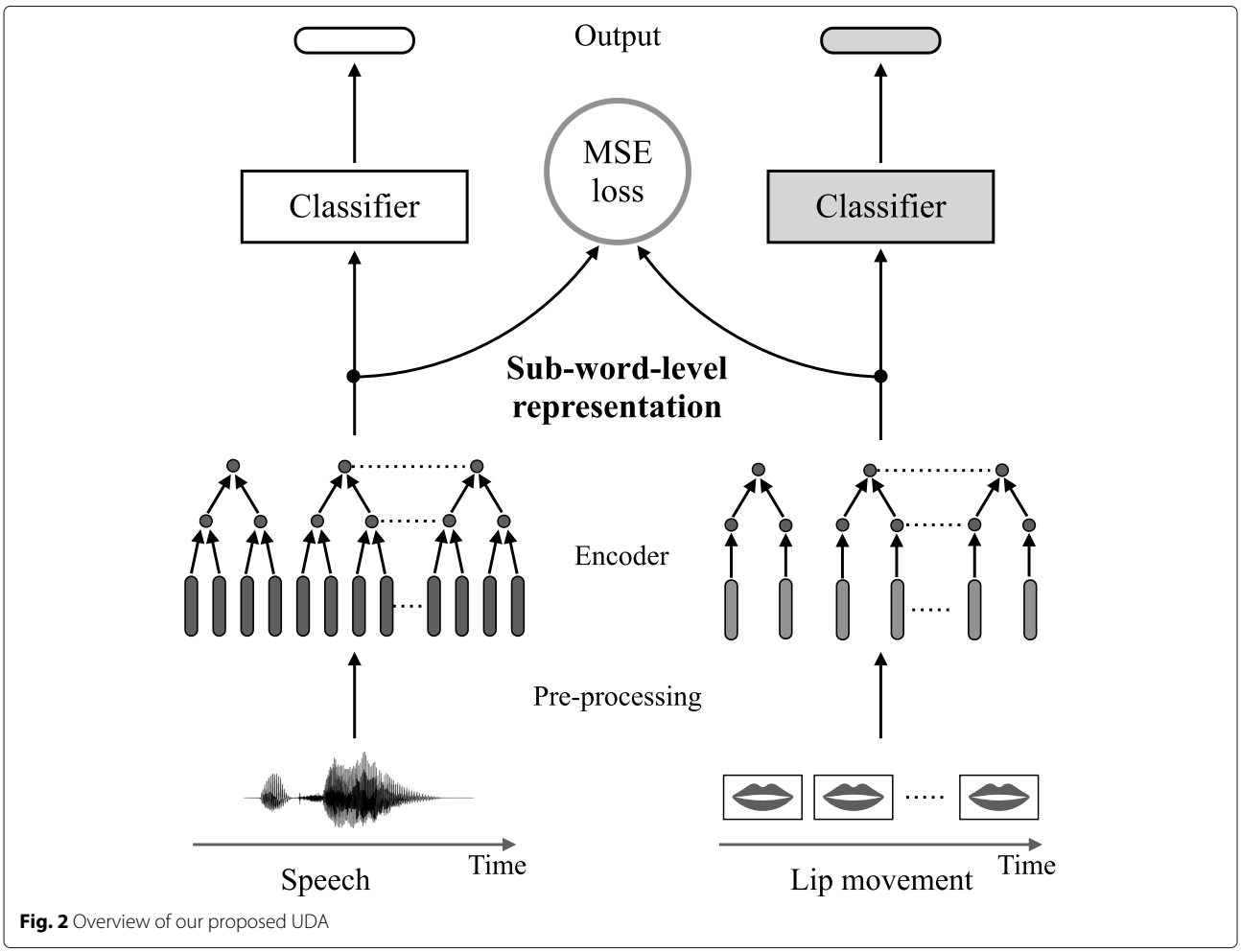

Fig. 2 Overview of our proposed UDA

3.3 Training procedure

Considering the source domain, let $\mathcal{D}_{\text{src}}$ be the joint distribution over sequences of audio features and visual features, and the corresponding label. $\mathcal{D}_{\text{trg}}$ is analogously defined for the target domain.

The first step is to train the models on the source domain. In this step, we expect that the hidden representation in the visual model is similar to that of the audio model. First, in advance, we train the audio model using the cross entropy loss with the correct label as follows:

$$\mathbb{E}_{\{\boldsymbol{v},\boldsymbol{a},\boldsymbol{y}\}\sim\mathcal{D}_{\text{src}}}[-\log(g_{\text{aud}}(\boldsymbol{y}|\boldsymbol{h}^{a}))], \tag{5}$$

where $\boldsymbol{v}$ is not used and $g_{\text{aud}}(\boldsymbol{y}|\boldsymbol{h}^{a})$ denotes the output probability of the label $\boldsymbol{y}$ estimated by the classifier of the audio model from the encoded feature $\boldsymbol{h}$. Then, we train the visual model using the KD loss and the cross entropy loss as follows:

$$\begin{aligned}&\mathcal{L}_{MSE}(\mathcal{D}_{\text{src}}) + \mathcal{L}_{CE}(\mathcal{D}_{\text{src}})\\ &= \mathcal{L}_{MSE}(\mathcal{D}_{\text{src}}) + \mathbb{E}_{\{\boldsymbol{v},\boldsymbol{a},\boldsymbol{y}\}\sim\mathcal{D}_{\text{src}}}[-\log(g_{\text{vis}}(\boldsymbol{y}|\boldsymbol{h}^{v}))],\end{aligned} \tag{6}$$

where $g_{\text{vis}}(\boldsymbol{y}|\boldsymbol{h}^{v})$ is the output probability estimated by the visual classifier.

Next, we adapt the visual model using the data of the unknown class based on the UDA scheme as described in Section 3.2. For more stable adaptation, we also use the data of the source domain. In addition to the loss in Eq. (4), we calculate losses for the source domain that has the correct label. This works as a regularization to prevent overfitting to the target distribution in the audio modality. Finally, our UDA approach for an unknown class minimizes the loss as follows:

$$\mathcal{L} = \mathcal{L}_{MSE}(\mathcal{D}_{\text{trg}}) + \alpha\mathcal{L}_{MSE}(\mathcal{D}_{\text{src}}) + (1-\alpha)\mathcal{L}_{CE}(\mathcal{D}_{\text{src}}), \tag{7}$$

where α indicates a weight parameter used to adapt the model stably, and we employ 0.5 in this paper. All parameters of the visual model are fine-tuned to minimize this loss function.

4 Experiments

4.1 Conditions

The proposed method was evaluated in a word recognition task on the lip reading in the wild (LRW) dataset [5]. LRW is a large-scale lip-reading dataset that consists of sounds and face images, where some works on AV-ASR or lip reading [31, 32] have been verified. All the videos are clipped to 29 frames (1.16 s) in length. Note that the length of each utterance is completely fixed.

LRW consists of up to 1000 utterances of 500 different words spoken by hundreds of different speakers. From the whole of the dataset, we picked out 800 utterances of 500 words (a total of 400,000 utterances) and divided them into several subsets, as shown in Fig. 3. We randomly divided 500 words into two sets of classes: the known class set of 400 words and the unknown class set of 100 words. For each of the 400 known words, we picked out (a) 500 utterances (a total of 200,000 utterances) and (b) another 50 utterances (a total of 20,000 utterances) as the training set of the source domain and the evaluation set of the target domain, respectively. For evaluating the UDA method, we used two different adaptation sets: the known class set and the unknown class set. The unknown class set (d) consisted of 250 utterances of 100 unknown words (a total of 25,000 utterances). For known class set (c), we randomly selected 100 words from 400 known words in order to match the condition with the unknown class set. Then, we created the known class set using 250 utterances of the selected 100 known words (a total of 25,000

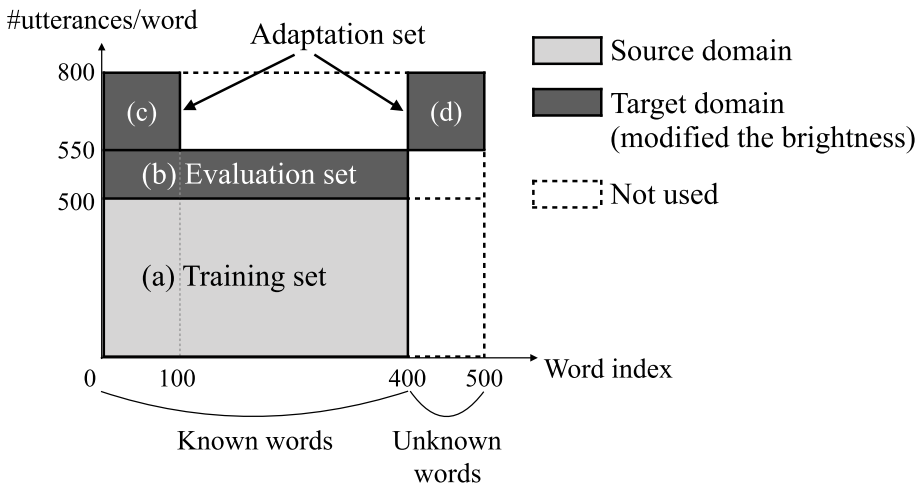

Fig. 3 Graphical representation of the division of the words, (a) source training set of the known words, (b) target evaluation set of the known words, (c) target adaptation set of the known words, and (d) target adaptation set of the unknown words

utterances) which were not used for either the training set or the evaluation set. The evaluation set and the two adaptation sets were in the target domain while the training set was in the source domain. For creating the target domain data, we changed the brightness of the image (no transformation was carried out on the sound signal of the video) because changes in brightness are one of the most likely situations in real environments (e.g., daytime and night, or a car navigation system when driving through a tunnel).

For the acoustic features, we calculated 40-dimensional log-mel filter bank features computed every 10 ms over a 25 ms window. Then, we stacked their delta and acceleration along the channel. The number of frames was 116. For the visual feature, the images are transformed to grayscale and resized to 112 ×112. The number of frames was 28. The encoder configuration is shown in Table 1. We used a pyramid structure that takes every two consecutive frames of the output from the previous layer without overlap as input. This structure allows the subsequent module to extract the relevant information from a smaller number of time steps. For the classifier, we use the three fully connected layers (4096 → 4096 → 400). We construct the individual model for the two modalities. The network was optimized using an Adam optimizer [33]. The batch size was 24, and the learning rate was set to 1e−4. When training models on the source domain, the number of epochs was 20 with early stopping. When adapting models to the target domain, the number of epochs was 10.

Our experiments were conducted using an Intel(R) Core(TM) i9-7900X CPU @ 3.30 GHz and single GeForce GTX 1080 Ti. Our proposed model took about 1.5 hours and 10 minutes per epoch for training and adaptation, respectively.

4.2 Results and discussion

First, we evaluated the performance of cross-modal KD on the training data of the source domain model. Here, we use the test data without modifying the brightness of the image and do not consider UDA. Table 2 shows the word recognition accuracy corresponding to each method. "Baseline" indicates the baseline model that was trained using the face image only. In our proposed method, we adopt a dimension reduction in the model (the 4th layer in Table 1) to calculate the KD loss efficiently. However, the dimension reduction is removed here for evaluating the baseline model because it degraded the recognition accuracy of the baseline model. From this table, when using 250 utterances per word to train the model, our proposed model achieved a relative improvement of 3.57% compared to the baseline model, despite the comparable performance when using 500 utterances per word. Typically, the audio data has more information for recognizing the speech and shows better recognition accuracy than the use of the visual data. This result shows that the output from the ASR model worked as a powerful teacher. We also assume that KD affects the regularization because we obtained more improvement with less training data.

Next, we confirmed the effectiveness of our proposed method for UDA. For the baseline adaptation, we updated parameters using two losses: the cross entropy loss of the source domain (the third term of Eq. (7)) and the pseudo label of the target domain estimated by the source model itself. Table 3 shows the word recognition accuracy corresponding to each method. Our proposed method outperformed conventional UDA in the setting that uses the known class. The relative improvements compared to no adaptation are 12.53% for the baseline method and 21.88% for our method, respectively. Moreover, our proposed method also improved the classification accuracy compared to no adaptation by relatively 21.48% even when we used unknown class adaptation data. These results

Table 1 Network architecture of the encoder

	Operation	
# Layer	**Audio model**	**Visual model**
Input	40×116×3	122×122 × 28×1
1	5×2 conv, 64, s(1,2)	3×3×1 conv, 48, p(1,1,0), 2×2 max-pool
2	5×2 conv, 128, s(1,2), 2×1 max-pool	3×3×2 conv, 96, s(1,1,2), 3×3 max-pool
3	5×2 conv, 256, 2×1 max-pool unfold along the time step	
4	128 dense*	

s(·) and p(·) indicate a stride size and a padding size, respectively
*A step-wise operation, which is applied for each time step independently
The activation function is ReLU

Table 2 Word recognition accuracy [%] for each method on the source domain

	# Utterance/word	
Model	**250**	**500**
Baseline	48.21	54.62
Proposed	50.06 (86.65)	55.07 (90.51)

#Utterance/word indicates the number of utterances per word used to train the model
The value in parentheses shows the accuracy of the audio model

Table 3 Word recognition accuracy [%] for each method

Eval. domain		**Baseline**	**Proposed**
Source		54.62	55.07
Target	No adap.	39.19	42.84
	Known class adap.	44.10	52.22
	Unknown class adap.	—	52.04

First column shows a domain of the evaluation data

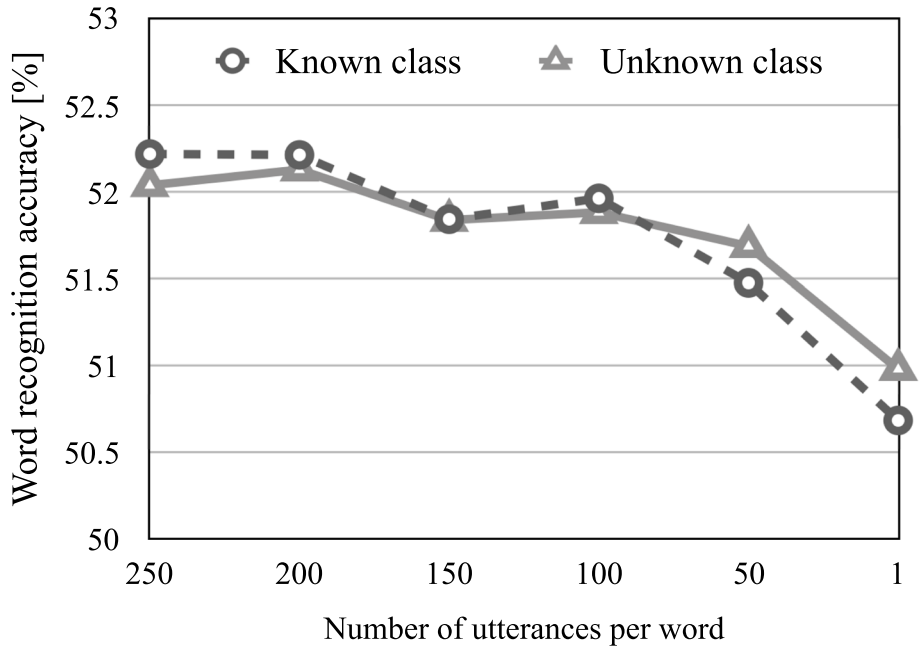

Fig. 4 The correlation between word recognition accuracies and the number of adaptation utterances. "Known class" and "Unknown class" indicate adaptation using the data of the known class and of the unknown class, respectively

show that the intermediate-layer-based KD approach can transfer the sub-class representation that does not depend on the class. Therefore, by using such representation as an objective of the adaptation, it is possible to use the unknown class data for UDA.

Moreover, we measured the performance of our UDA approach as a function of the number of adaptation utterances. As shown in Fig. 4, we observed that the accuracy decreases as the number of adaptation utterances decreases. We can see that the accuracies are saturated when using about 200 utterances for adaptation. Moreover, even when we use a smaller amount of the adaptation data, our method can adapt the model more effectively than the baseline method using all of the adaptation data (the fourth row in Table 3). These results demonstrate that our method can achieve stable and effective adaptation for UDA using the data of the unknown class.

Finally, we calculated the real time factor (RTF) that is the ratio of the recognition response to the utterance duration. Generally, RTF <1 is required for real-time scenarios. Here, decoding was performed on an Intel(R) Core(TM) i9-7900X CPU @ 3.30 GHz. The RTF of our system was 1.16. We consider that using a more efficient network architecture, such as MobileNet [34, 35], could improve the RTF while maintaining the performance.

4.3 Changing the division of the known/unknown words

In the experiments mentioned above, we used the fixed split for the known/unknown words. To evaluate the robustness of the variety of division pattern of the known/unknown words, we conducted 5-fold cross-validation for our proposed method. For this purpose, we split 500 words in LRW into 5 consecutive folds. Then, we used 100 words as the unknown class and the remaining 400 words as the known class. Table 4 shows the word recognition accuracy corresponding to each fold. The rightmost column in the table shows the mean value and the standard deviation. Our proposed method had a small standard deviation. This means that our method has high robustness for the selection of the words.

4.4 Noisy audio

To demonstrate the potential of our proposed method, we conducted the experiments in a more realistic scenario. For this purpose, we introduced acoustic noise for the audio in addition to brightness for the image during adaptation. White noise was added to audio signals, and their SNR was set to 30dB, 20dB, 10dB, and 0dB. As

Table 4 Word recognition accuracy [%] for the 5-fold cross-validation

	5-folds					
Method	**1st**	**2nd**	**3rd**	**4th**	**5th**	**mean**
Known class adap.	52.22	52.79	51.15	53.95	52.68	52.56±1.02
Unknown class adap.	52.04	52.71	50.91	53.87	52.75	52.46±1.09

The results of the 1st fold correspond to those in Table 3

Table 5 Word recognition accuracy [%] corresponding to each SNR

Method	clean	30dB	20dB	10dB	0dB
Known class adap.	52.22	47.50	47.55	47.33	47.13
Unknown class adap.	52.04	47.18	47.21	47.10	46.75

The results of "clean" correspond to those in Table 3

shown in Table 5, although the performance of our proposed method hardly varies among different SNRs (less than 1%), the use of the noisy audio signal significantly degraded the adaptation performance compared to using a clean audio signal.

By comparing the results of "clean" and "30dB" in Table 5, we see that the recognition accuracy greatly degraded even though "30dB" was a small noise condition. In order to analyze these results, we measured how greatly the hidden representation of the audio signal $\boldsymbol{h}^a$, which is used as a teacher in our proposed cross-modal KD for UDA (see Eq. 4), is distorted by noise under each condition. For this measurement, we calculated the SNR under the hidden representation space as follows:

$$\mathrm{SNR} = 10\log_{10}\frac{||\boldsymbol{h}^a_{\mathrm{clean}}||_2^2}{||\boldsymbol{h}^a_{\mathrm{noisy}} - \boldsymbol{h}^a_{\mathrm{clean}}||_2^2}, \tag{8}$$

where $\boldsymbol{h}^a_{\mathrm{clean}}$ and $\boldsymbol{h}^a_{\mathrm{noisy}}$ denote the hidden representations $\boldsymbol{h}^a$ obtained under clean and noisy (SNR = 30, 20, 10, 0dB) conditions, respectively. Table 6 shows the SNR of $\boldsymbol{h}^a$ for each SNR of the input audio signal. As shown in this table, even when the SNR of the input audio signal was 30dB, the SNR of the hidden representation degraded to 14.14dB. Because this distorted hidden representation was used as a teacher in our proposed cross-modal KD, this result means that the proposed method is sensitive to the noise in the input audio signal. One possible reason for this sensitivity is that the audio model was trained using clean speech data and overfitted to the clean condition. Therefore, this degradation might be reduced if we use noisy audio data to train the noise-robust audio model. Nevertheless, the performance of our proposed system using the noisy audio signal still outperformed the baseline system (44.10, Known class adap. in Table 3) and the proposed system without adaptation (42.84 in Table 3) which do not use the audio signal.

Table 6 SNR of the hidden representation $\boldsymbol{h}^a$ for each SNR of the input audio signal

SNR of audio signal	30dB	20dB	10dB	0dB
SNR of $\boldsymbol{h}^a$	14.14dB	7.23dB	2.69dB	-0.14dB

5 Conclusion

In this paper, we proposed the intermediate-layer-based KD approach for UDA, which can effectively transfer the knowledge of the ASR model to the lip-reading model. Our method allows us to use the data of the unknown class to adapt the model from the source domain to the target domain. Experimental results show that our proposed method can adapt the model effectively regardless of whether the class of the adaptation data is known or unknown.

We used a simple network architecture based on stacked convolution layers because we assume an isolated word recognition task. In order to extend our approach for a continuous speech recognition task (i.e., sentence recognition task), we will investigate the use of recurrent neural network-based models which are suitable for this task, such as LipNet [17], in the future. In addition, we will demonstrate the effectiveness of our method in more complex transformations or more realistic environments. Our proposed method can use the audio-only database because the ASR model and the lip reading model are trained separately. Therefore, we will further investigate the combination with large audio databases. Our future work will also include the further investigation of its potential, focusing particularly on multi-modal tasks.

Abbreviations

ASR: Automatic speech recognition; AV: Audio-visual; DA: Domain adaptation; GAN: Generative adversarial network; KD: Knowledge distillation; LRW: Lip reading in the wild; OOV: Out of vocabulary; RTF: Real time factor; UDA: Unsupervised domain adaptation

Acknowledgements

Not applicable.

Authors' contributions

The first author mainly performed the experiments and wrote the paper, and the other authors reviewed and edited the manuscript. All of the authors discussed the final results. All of the authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

All data used in this study are included in the lip reading in the wild (LRW) dataset [5].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Graduate School of System Informatics, Kobe University, Kobe, Japan. [2] Information Technology R&D Center, Mitsubishi Electric Corporation, Ofuna, Japan.

Received: 8 July 2021 Accepted: 11 November 2021
Published online: 11 December 2021

References

1. H. McGurk, J. MacDonald, Hearing lips and seeing voices. Nature. **264**, 746–748 (1976)
2. M. J. Tomlinson, M. J. Russell, N. M. Brooke, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Integrating audio and visual information to provide highly robust speech recognition, (1996), pp. 821–824
3. A. Verma, T. Faruquie, C. Neti, S. Basu, A. Senior, in *Proc. IEEE Automatic Speech Recognition and Understanding (ASRU)*. Late integration in audio-visual continuous speech recognition, vol. 1, (1999), pp. 71–74
4. K. Palecek, J. Chaloupka, in *Proc. International Conference on Telecommunications and Signal Processing (TSP)*. Audio-visual speech recognition in noisy audio environments, (2013), pp. 484–487
5. J. S. Chung, A. Zisserman, in *Proc. Asian Conference on Computer Vision (ACCV)*. Lip reading in the wild, (2016), pp. 87–103
6. J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Lip reading sentences in the wild, (2017), pp. 3444–3453
7. J. Yu, S. Zhang, J. Wu, S. Ghorbani, B. Wu, S. Kang, S. Liu, X. Liu, H. Meng, D. Yu, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Audio-visual recognition of overlapped speech for the LRS2 dataset, (2020), pp. 6984–6988
8. Y. Ganin, V. S. Lempitsky, in *Proc. International Conference on Machine Learning (ICML)*. Unsupervised domain adaptation by backpropagation, (2015), pp. 1180–1189
9. M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, W. Li, in *Proc. European Conference on Computer Vision (ECCV)*. Deep reconstruction-classification networks for unsupervised domain adaptation, vol. 9908, (2016), pp. 597–613
10. K. Saito, K. Watanabe, Y. Ushiku, T. Harada, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Maximum classifier discrepancy for unsupervised domain adaptation, (2018), pp. 3723–3732
11. P. P. Busto, J. Gall, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Open set domain adaptation, (2017), pp. 754–763
12. G. Hinton, O. Vinyals, J. Dean, in *Proc. NIPS Deep Learning Workshop*. Distilling the knowledge in a neural network, (2014)
13. T. Asami, R. Masumura, Y. Yamaguchi, H. Masataki, Y. Aono, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Domain adaptation of DNN acoustic models using knowledge distillation, (2017), pp. 5185–5189
14. G. Chen, W. Choi, X. Yu, T. X. Han, M. Chandraker, in *NIPS*. Learning efficient object detection models with knowledge distillation, (2017), pp. 742–751
15. W. Li, S. Wang, M. Lei, S. M. Siniscalchi, C. H. Lee, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Improving audio-visual speech recognition performance with cross-modal student-teacher training, (2019), pp. 6560–6564
16. H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, K. Takeda, in *Proc. ISCA Interspeech*. Integration of deep bottleneck features for audio-visual speech recognition, (2015), pp. 563–567
17. Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas, LipNet: Sentence-level lipreading (2016). arXiv preprint arXiv:1611.01599
18. A. Graves, S. Fernández, F. J. Gomez, J. Schmidhuber, in *Proc. International Conference on Machine Learning (ICML)*. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, (2006), pp. 369–376
19. A. Koumparoulis, G. Potamianos, in *Proc. ISCA Interspeech*. MobiLipNet: Resource-efficient deep learning based lipreading, (2019), pp. 2763–2767
20. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, in *NIPS*. Generative adversarial nets, (2014), pp. 2672–2680
21. M. Wand, J. Schmidhuber, in *Proc. ISCA Interspeech*. Improving speaker-independent lipreading with domain-adversarial training, (2017), pp. 3662–3666
22. D. A. B. Oliveira, A. B. Mattos, E. D. S. Morais, in *Proc. European Conference on Computer Vision (ECCV) Workshops*. Improving viseme recognition using GAN-based frontal view mapping, (2018), pp. 2148–2155
23. S. Gupta, J. Hoffman, J. Malik, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Cross modal distillation for supervision transfer, (2016), pp. 2827–2836
24. T. Afouras, J. S. Chung, A. Zisserman, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ASR is all you need: Cross-modal distillation for lip reading, (2020), pp. 2143–2147
25. Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, M. Song, in *Proc. The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*. Hearing lips: Improving lip reading by distilling speech recognizers, (2020), pp. 6917–6924
26. E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: Maximizing for domain invariance (2014). arXiv preprint arXiv:1412.3474
27. E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Adversarial discriminative domain adaptation, (2017), pp. 2962–2971
28. R. Shu, H. H. Bui, H. Narui, S. Ermon, in *Proc. International Conference on Learning Representations (ICLR)*. A DIRT-T approach to unsupervised domain adaptation, (2018)
29. K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, M. Chandraker, in *Proc. IEEE International Conference on Computer Vision (ICCV)*. Unsupervised domain adaptation for face recognition in unlabeled videos, (2017), pp. 5917–5925
30. A. Mouchtaris, J. V. der Spiegel, P. Mueller, Nonparallel training for voice conversion based on a parameter adaptation approach. IEEE Trans. Audio Speech Lang. Process. **14**(3), 952–963 (2006)
31. S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, M. Pantic, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. End-to-end audiovisual speech recognition, (2018), pp. 6548–6552
32. T. Stafylakis, G. Tzimiropoulos, in *Proc. ISCA Interspeech*. Combining residual networks with LSTMs for lipreading, (2017), pp. 3652–3656
33. D. P. Kingma, J. Ba, in *Proc. International Conference on Learning Representations (ICLR)*. Adam: A method for stochastic optimization, (2015)
34. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient convolutional neural networks for mobile vision applications (2017). arXiv preprint arXiv:1704.04861
35. M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. MobileNetV2: Inverted residuals and linear bottlenecks, (2018), pp. 4510–4520

Publisher's Note