Speech Recognition using Deep Canonical Correlation Analysis in Noisy Environments

Shinnosuke Isobe, Satoshi Tamura and Satoru Hayamizu

Gifu University, Gifu, Japan

Keywords: Speech Recognition, Audio-visual Processing, Canonical Correlation Analysis, Noise Robustness, Data Augmentation, Deep Learning.

Abstract: In this paper, we propose a method to improve the accuracy of speech recognition in noisy environments by utilizing Deep Canonical Correlation Analysis (DCCA). DCCA generates projections from two modalities into one common space, so that the correlation of projected vectors could be maximized. Our idea is to employ DCCA techniques with audio and visual modalities to enhance the robustness of Automatic Speech Recognition (ASR); A) noisy audio features can be recovered by clean visual features, and B) an ASR model can be trained using audio and visual features, as data augmentation. We evaluated our method using an audio-visual corpus CENSREC-1-AV and a noise database DEMAND. Compared to conventional ASR and feature-fusion-based audio-visual speech recognition, our DCCA-based recognizers achieved better performance. In addition, experimental results shows that utilizing DCCA enables us to get better results in various noisy environments, thanks to the visual modality. Furthermore, it is found that DCCA can be used as a data augmentation scheme if only a few training data are available, by incorporating visual DCCA features to build an audio-only ASR model, in addition to audio DCCA features.

1 INTRODUCTION

Recently, ASR has become widely spread, and is used in various scenes 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 such the ASR schemes applicable in noisy environments is multimodal speech recognition; most multimodal ASR methods employ speech signals and lip images, which are not affected by audio noise (K.Noda et al., 2015; W.Feng et al., 2017a);

In recent high-performance multimodal ASR systems, visual features are extracted from lip images using Convolutional Neural Networks (CNNs). In order to build CNN models, generally, a well-labeled large-size training dataset is required. However, collecting training data in real environments and annotating the recorded data are quite hard and expensive. In the deep-learning fields, to compensate this, data augmentation is often applied, which artificially generates new training data from the original training data (T.Tran et al., 2017; H.I.Fawaz et al., 2018). For example, additional training data are obtained by overlapping audio noise in speech recognition (A.Graves et al., 2013), and in computer vision new image data are generated by rotating or flipping original images (T.DeVries and G.W.Taylor, 2017).

In this paper, we propose a new data augmentation approach for ASR based on DCCA (G.Andrew et al., 2013), to improve the accuracy in noisy conditions. The DCCA method basically enhances the correlation between lip image features and corresponding clean audio features as well as noisy ones contaminated by audio noise. It is well known that clean audio features have much more information than visual features regarding speech recognition. According to this fact, we have proposed a lipreading scheme enhanced by audio information with DCCA. On the other hand, speech signals recorded in real environments often consist of uncertain or degraded information, in which visual information is more useful. Therefore, there is a room to improve audio features by employing the DCCA technique with visual information. In addition, we can apply our proposed approach to conduct the data augmentation, particularly in the case we suffer from the lack of audio training

Isobe, S., Tamura, S. and Hayamizu, S.
Speech Recognition using Deep Canonical Correlation Analysis in Noisy Environments.
DOI: 10.5220/0010268200630070
In *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2021)*, pages 63-70
ISBN: 978-989-758-486-2

data for Deep Neural Networks (DNNs); since DCCA projects features in two modalities into one common feature space, we can easily convert visual features into audio ones as the augmentation data.

In order to clarify the effectiveness of our scheme, we have two experiments using audio-visual and noise corpora. In the first experiment, we evaluated and compared some ASR methods to show the advantage of DCCA. The second experiment tried to investigate how the DCCA-based data augmentation might work for ASR, incorporating the visual modality.

The rest of this paper is organized as follows. At first, we describe related works in Section 2. Next, Section 3 briefly introduces CCA and DCCA. ASR models used in the experiments are explained in Section 4. Experimental setup, results, and discussion are described in Section 5. Finally Section 6 concludes this paper.

2 RELATED WORK

Recently, many researchers have proposed deep-learning-based audio-visual ASR and lipreading schemes. A large-scale audio-visual speech recognition system based on a Recurrent-Neural-Network Transducer (RNN-T) architecture was built in (T.Makino et al., 2019). The high performance in the LRS3-TED set had been confirmed by using this system. In (P.Zhou et al., 2019), the authors proposed a multimodal attention-based method for audio-visual speech recognition which could automatically learn the fused representation from both modalities based on their importance. This model employed sequence-to-sequence architectures, and they confirmed high recognition performance under both clean and noisy conditions. An audio-visual automatic speech recognition system using a transformer-based architecture was proposed (G.Paraskevopoulos et al., 2020). Experimental results show that on the How2 dataset, relatively the system improved word error rate over sub-word prediction models.

As research works much more related to this paper, (Joze et al., 2020) introduced the transducer technique to fuse CNNs for two modalities. They investigated the effectiveness of their approach in several tasks, such as audio-visual speech enhancement. There is another work which tried to incorporate audio and visual modality (W.Feng et al., 2017b). They concatenated audio and visual features obtained from DNNs and put them into a multimodal layer to get the final decision. In (A.Renduchintala et al., 2018), the authors investigated a new end-to-end architecture for ASR that can be trained using symbolic input in addition to the traditional acoustic input system. They called the architecture multimodal data augmentation.

Some of authors of this paper have reported a multimodal data augmentation scheme based on DCCA (M.Shimonishi et al., 2019). In this method, audio features that were enhanced correlation with image features by DCCA were used as augmentation data in lipreading. After getting audio and visual features, DCCA was applied to obtain audio DCCA and visual DCCA features. Experimental results shows that the model trained using both DCCA features achieved better performance than those obtained from only image DCCA features and audio DCCA features, respectively.

3 DEEP CANONICAL CORRELATION ANALYSIS

3.1 CCA

CCA is a method to make projections from two modalities so that the correlation between transformed vectors should be maximized. When observing a pair of two vectors $\boldsymbol{x}$, $\boldsymbol{y}$, we can linearly transform them using CCA as follows, with the parameter vectors $\boldsymbol{a}$, $\boldsymbol{b}$:

$$u(\boldsymbol{x}) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \quad , \quad v(\boldsymbol{y}) = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} \tag{1}$$

CCA tries to find the parameters $\boldsymbol{a}^*$, $\boldsymbol{b}^*$ such that maximize the following correlation coefficient:

$$\begin{aligned} (\boldsymbol{a}^*, \boldsymbol{b}^*) &= \underset{\boldsymbol{a},\boldsymbol{b}}{\operatorname{argmax}} \ \operatorname{corr}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}) \\ &= \underset{\boldsymbol{a},\boldsymbol{b}}{\operatorname{argmax}} \ \frac{\boldsymbol{a}^{\mathrm{T}} V_{\boldsymbol{xy}} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^{\mathrm{T}} V_{\boldsymbol{xx}} \boldsymbol{a}} \sqrt{\boldsymbol{b}^{\mathrm{T}} V_{\boldsymbol{yy}} \boldsymbol{b}}} \end{aligned} \tag{2}$$

where $V_{\boldsymbol{xx}}, V_{\boldsymbol{yy}}, V_{\boldsymbol{xy}}$ are variance-covariance matrices. In Equation (2), we can regard standard deviation of denominator as 1. Finally, we can estimate the CCA parameters by solving the following constrained quadratic function maximization problem:

$$\max_{\boldsymbol{a},\boldsymbol{b}} \boldsymbol{a}^{\mathrm{T}} V_{\boldsymbol{xy}} \boldsymbol{b} \quad \text{s.t. } (\boldsymbol{a}^{\mathrm{T}} V_{\boldsymbol{xx}} \boldsymbol{a}) = (\boldsymbol{b}^{\mathrm{T}} V_{\boldsymbol{yy}} \boldsymbol{b}) = 1. \tag{3}$$

3.2 DCCA

Kernel-based CCA (KCCA) (R.Arora and K.Livescu, 2012) and deep learning-based CCA (DCCA) are non-linear CCA approaches. In this paper, we adopt DCCA. An overview of the DCCA is shown in Fig 1. We assume two DNNs f_1, f_2 in this paper each having two hidden layers. Vectors $\boldsymbol{x}, \boldsymbol{y}$ observed in two

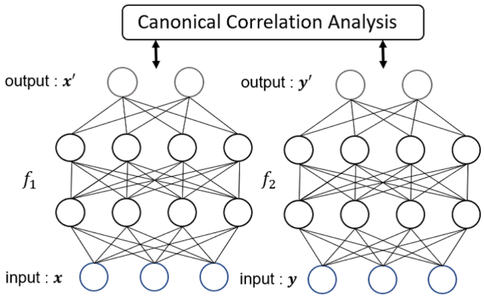


Figure 1: An overview of deep canonical correlation analysis, consisting of two DNNs, f_1 and f_2.

modalities respectively, are converted to $\boldsymbol{x}^{'},\boldsymbol{y}^{'}$ using f_1, f_2:

$$\boldsymbol{x}^{'} = f_1(\boldsymbol{x}) \quad , \quad \boldsymbol{y}^{'} = f_2(\boldsymbol{y}) \tag{4}$$

We then apply the linear CCA to $\boldsymbol{x}^{'}$ and $\boldsymbol{y}^{'}$ as:

$$S = \mathrm{corr}(\boldsymbol{x}^{'},\boldsymbol{y}^{'}). \tag{5}$$

In DCCA, we try to find parameters that maximize S in Equation (5). Assume $\boldsymbol{x}^{'}$ and $\boldsymbol{y}^{'}$ be standardized; $\bar{\boldsymbol{x}}^{'}$ and $\bar{\boldsymbol{y}}^{'}$ have an average of $\mathbf{0}$ and a variance of $\mathbf{1}$. The variance-covariance matrices of $\bar{\boldsymbol{x}}^{'}$, $\bar{\boldsymbol{y}}^{'}$ are expressed as follows:

$$\begin{aligned}\Sigma_{11} &= \frac{1}{N-1}\bar{\boldsymbol{x}}^{'}\bar{\boldsymbol{x}}^{'T} + r_1 I \\ \Sigma_{22} &= \frac{1}{N-1}\bar{\boldsymbol{y}}^{'}\bar{\boldsymbol{y}}^{'T} + r_2 I \\ \Sigma_{12} &= \frac{1}{N-1}\bar{\boldsymbol{x}}^{'}\bar{\boldsymbol{y}}^{'T},\end{aligned} \tag{6}$$

where $r_1 I$ and $r_2 I$ are regularization terms, r_1, r_2 (r_1, r_2>0) are regularization parameters, and N is the number of vectors. Here, the total correlation of the top k components of $\boldsymbol{x}^{'}$ and $\boldsymbol{y}^{'}$ is the sum of the top k singular values of the matrix $A = \hat{\Sigma}_{11}^{-1/2}\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1/2}$. If the number of output layer units is chosen as k, then this exactly corresponds to the matrix trace norm of A:

$$S = tr(A) = tr(A^T A)^{\frac{1}{2}} \tag{7}$$

Now the DCCA parameter update is performed based on the gradient descent, using Equation (7).

4 MODEL

In this paper, we adopt four models as shown in the Table 1. For comparison, an audio-only speech recognition model (ASR) and an end-to-end early-fusion multimodal speech recognition model (E2E-MM) are chosen. As our method, two models are also proposed in which DCCA is applied to audio-only ASR (A-DCCA) and audio-visual ASR (AV-DCCA), respectively.

Table 1: Four models used in this paper.

ASR	Audio-only speech recognition
E2E-MM	End-to-End early-fusion multimodal speech recognition
A-DCCA	Audio DCCA recognition
AV-DCCA	Audio-Visual DCCA recognition

Table 2: An architecture of ASR model.

Layer / Act.func.	Ch.	Kernel size	Output shape
Input	-	-	(39,33, 1)
Conv2D + BN / ReLU	32	(3,3)	(39,33,32)
MaxPooling2D	-	(2,2)	(20,17,32)
Conv2D + BN / ReLU	64	(3,3)	(20,17,64)
MaxPooling2D	-	(2,2)	(10, 9,64)
Conv2D + BN / ReLU	96	(3,3)	(10, 9,96)
GlobalMaxPooling2D	-	-	(96)
FC / ReLU	-	-	(40)
FC / softmax	-	-	(12)

* BN = Batch Normalization

4.1 ASR

The flow of the ASR model is shown in Fig 2 (A). We extract 13 Mel-Frequency Cepstral 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. MFCC is the most commonly used feature in the speech recognition field, in addition to ΔMFCC and ΔΔMFCC that are first and second derivatives, respectively. To conduct speech recognition in the same condition, in this paper, the current MFCC vector as well as previous 16 vectors and following 16 ones are concatenated into a 39×33 feature matrix, which is classified by 2D-CNN model, shown in Table 2.

4.2 E2E-MM

The flow of the E2E-MM model is shown in Fig 2 (B). To consecutive 11 image frames corresponding to 33 audio frames, a 3D-CNN model is applied to obtain a 40-dimensional visual feature. A 40-dimensional audio vector is also computed on the same manner except that a 2D-CNN model is adopted. Thereafter, we generate an audio-visual vector consisting of those vectors. Finally, recognition is performed by three Fully Connected (FC) layers. Details of the E2E-MM model are indicated in Tables 3, 4 and 5.

Table 3: An architecture of E2E-MM model (Visual).

Layer / Act.func.	Ch.	Kernel size	Output shape
Input	-	-	(11,48,64, 1)
Conv3D + BN / ReLU	8	(3,3,3)	(11,48,64, 8)
Conv3D + BN / ReLU	8	(3,3,3)	(11,48,64, 8)
Conv3D + BN / ReLU	8	(3,3,3)	(11,48,64, 8)
MaxPooling3D	-	(2,2,2)	(6,24,32, 8)
Conv3D + BN / ReLU	16	(3,3,3)	(6,24,32,16)
Conv3D + BN / ReLU	16	(3,3,3)	(6,24,32,16)
Conv3D + BN / ReLU	16	(3,3,3)	(6,24,32,16)
MaxPooling3D	-	(2,2,2)	(3,12,16,16)
Conv3D + BN / ReLU	32	(3,3,3)	(3,12,16,32)
Conv3D + BN / ReLU	32	(3,3,3)	(3,12,16,32)
Conv3D + BN / ReLU	32	(3,3,3)	(3,12,16,32)
MaxPooling3D	-	(2,2,2)	(2, 6, 8,32)
Conv3D + BN / ReLU	64	(3,3,3)	(2, 6, 8,64)
Conv3D + BN / ReLU	64	(3,3,3)	(2, 6, 8,64)
Conv3D + BN / ReLU	64	(3,3,3)	(2, 6, 8,64)
GlobalMaxPooling3D	-	-	(64)
FC / ReLU	-	-	(40)

* BN = Batch Normalization

Table 4: An architecture of E2E-MM model (Audio).

Layer / Act.func.	Ch.	Kernel size	Output shape
Input	-	-	(39,33, 1)
Conv2D + BN / ReLU	32	(3,3)	(39,33,32)
MaxPooling2D	-	(2,2)	(20,17,32)
Conv2D + BN / ReLU	64	(3,3)	(20,17,64)
MaxPooling2D	-	(2,2)	(10, 9,64)
Conv2D + BN / ReLU	96	(3,3)	(10, 9,96)
GlobalMaxPooling2D	-	-	(96)
FC / ReLU	-	-	(40)

* BN = Batch Normalization

4.3 DCCA Model

Both proposed models, A-DCCA and AV-DCCA, consist of the feature extraction part and the recognition part, described below.

4.3.1 Feature Extraction

The flow of feature extraction is shown in Fig 3. At first, we conduct preliminary recognition in the image and audio modalities, respectively, to build their CNN models. The structure of the audio CNN model is the same as the ASR model, and the image model has additional FC layer at the end of Table 3, to classify lip images. Second, 40-dimensional audio and image feature vectors are calculated as same as E2E-MM. By applying DCCA, we finally generate visual and audio DCCA features, each including 40 coefficients. Refer to Tables 3 and 4 for CNN architectures. Note that the architecture of DCCA is the same as Fig 1, in which there are 1,600 units on each hidden layer.

Table 5: An architecture of E2E-MM model (AV).

Layer / Act.func.	Output shape
Input	(40),(40)
Merge	(80)
FC / ReLU	(160)
Dropout(0.5)	-
FC / ReLU	(160)
Dropout(0.5)	-
FC / softmax	(12)

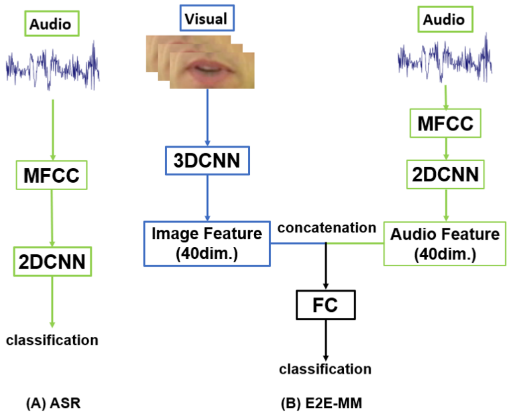


Figure 2: An overview of ASR model (A) and E2E-MM model (B).

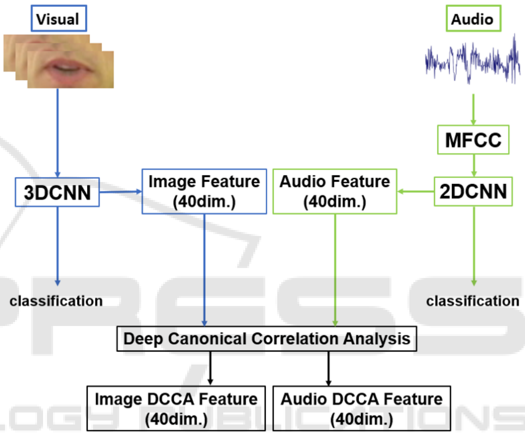
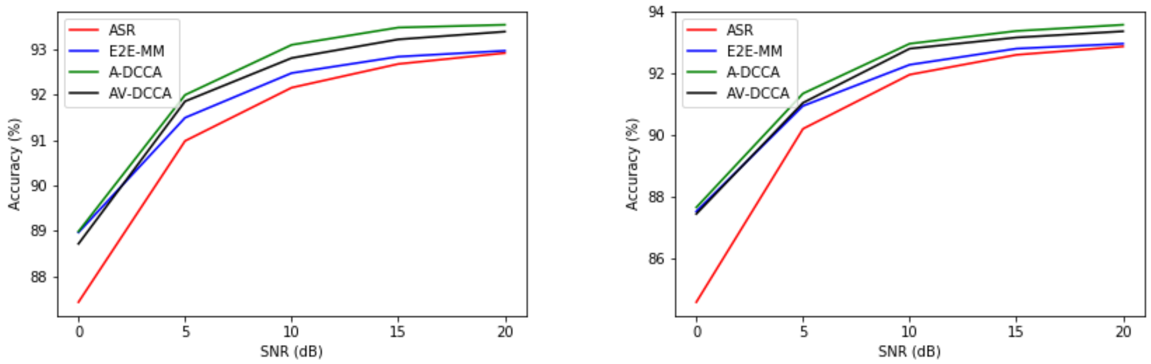


Figure 3: Feature extraction for DCCA models.

4.3.2 Recognition

For DCCA feature vectors, a recognition model consisting of three FC layers is prepared, as shown in Table 6. The A-DCCA model takes audio DCCA features as input, while the AV-DCCA model accepts not only audio DCCA features but also visual DCCA features.



(a) Cityroad

(b) Expressway

Figure 4: Frame-level Recognition Accuracy [%] in Experiment A.

Table 6: An architecture of recognition model.

Layer / Act.func.	Output shape
Input	(40)
FC / ReLU	(160)
Dropout(0.25)	-
FC / ReLU	(160)
Dropout(0.25)	-
FC / softmax	(12)

Table 7: Digit pronunciation of CENSREC-1-AV.

Word	Phonemes	Word	Phonemes
1 (one)	/ichi/	7 (seven)	/nana/
2 (two)	/ni/	8 (eight)	/hachi/
3 (three)	/saN/	9 (nine)	/kyu/
4 (four)	/yoN/	0 (oh)	/maru/
5 (five)	/go/	Z (zero)	/zero/
6 (six)	/roku/		

5 EXPERIMENTS

5.1 Database

5.1.1 CENSREC-1-AV

We chose a Japanese audio-visual corpus for noisy multimodal speech recognition, CENSREC-1-AV (S.Tamura et al., 2010). CENSREC-1-AV is providing training data and testing data. The training data set includes 3,234 utterances spoken by 20 female and 22 male subjects. The testing data set includes 1,963 utterances spoken by 26 female and 25 male subjects. We further split the training data into two sets: 90% for DNN training and 10% for validation. Each utterance in this database consists of 1-7 digits; one subject in the training set uttered 77 utterances while one speaker spoke 38-39 utterances in the testing set. Table 7 indicates Japanese digit pronounces in CENSREC-1-AV. To train recognition models and to evaluate them in different noise conditions, not only clean data but also noisy data were prepared for our experiments; interior car noises recorded on cityroads and expressways were adopted.

5.1.2 DEMAND

We selected another database DEMAND (J.Thiemann and N.Ito, 2013) as a noise corpus. This corpus consists of six primary categories, each which has three environments, respectively. Four of those primary categories are for closed spaces: Domestic, Office, Public, and Transportation. And the remaining two categories were recorded outdoors: Nature and Street.

5.2 Experimental Setup

We evaluated model performance by frame-level accuracy:

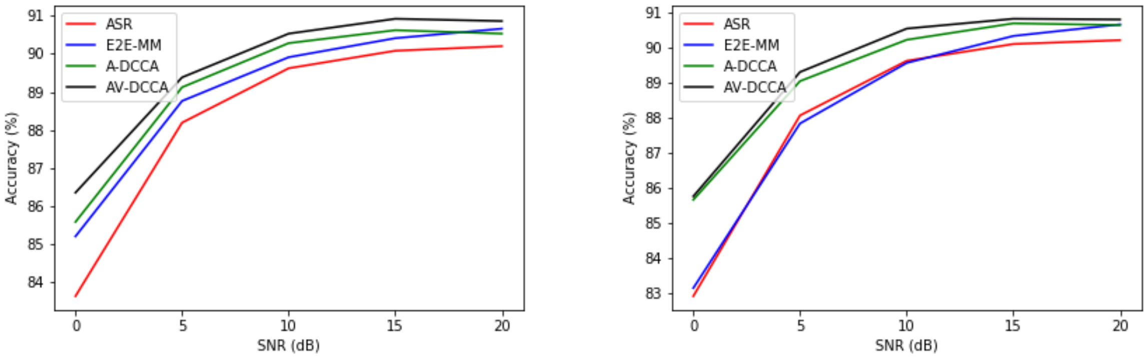
$$\text{Accuracy} = \frac{H}{N} \times 100\ [\%] \qquad (8)$$

where H and N are the number of correctly recognized frames and the total number of frames, respectively. Since DNN-based model performance slightly varies depending on the probabilistic gradient descend algorithm, we repeated the same experiment 5 times, and the mean of accuracy is calculated.

In terms of DNN hyperparameters, we chose the cross-entropy loss function with the Adam optimizer. The batch size was set to 32.

5.2.1 Experiment A: The Effectiveness of using the DCCA

We verified the effectiveness of employing DCCA for ASR. The four models described in Section 4 were used for recognition. Regarding the training data, utterances of 12 speakers were kept clean, while noise was added to the rest; among five kinds of noises in DEMAND, each noise was overlapped to speech data



(a) Cityroad

(b) Expressway

Figure 5: Frame-level Recognition Accuracy [%] in Experiment B.

Table 8: Data Specification in Experiment A.

		Noise	#spkr	#data
Train	Audio	Clean	12	68,412
		Domestic	6	32,854
		Office	6	32,848
		Public	6	32,498
		Transportation	6	35,849
		Nature	6	33,409
	Visual	Clean	42	235,870
Test	Audio	Cityroad	51	145,647
		Expressway	51	145,647
	Visual	Clean	51	145,647

Table 9: Data Specification in Experiment B.

		Noise	#spkr	#data
Train	Audio	Clean	2	12,605
		Domestic	2	11,590
		Office	2	13,131
		Public	2	11,118
		Transportation	2	11,120
		Nature	2	11,516
	Visual	Clean	12	71,080
Test	Audio	Cityroad	51	145,647
		Expressway	51	145,647
	Visual	Clean	51	145,647

of six speakers, at either of 0dB, 5dB, 10dB, 15dB, and 20dB. For the test data, two types of noises in CENSREC-1-AV were added with either of the five SNRs to generate 10 different audio data. In both cases, we did not use visual noises. Table 8 summarizes the data used in this experiment. In Table 8, #spkr represents the number of speakers, and #data represents the number of framed data.

5.2.2 Experiment B: DCCA-based Data Augmentation

We also evaluated the effectiveness of using DCCA when the amount of training data is not enough to learn DNN models. In other words, we investigated whether visual DCCA features can be used as augmentation data for audio-only ASR. As shown in Table 9, we built a small training data set including clean and noisy speeches and corresponding visual data, by randomly choosing 12 speakers from the original training set. The other setup was the same as Experiment A.

5.3 Result and Discussion

5.3.1 Experiment A

Fig 4 shows recognition results using each model under the noises. For comparison, the accuracy of lipreading only using visual information is 74.02%. We firstly compare ASR and E2E-MM. E2E-MM has better accuracy than ASR. As we mentioned, speech recognition is generally better than lipreading. Nevertheless, multimodal speech recognition that employs lip image features and audio features is better than speech recognition in noisy environments. That means visual information contributes to the improvement of speech recognition.

Next, we compare those with A-DCCA and AV-DCCA. Among the models, A-DCCA has the best accuracy in all the SNRs, followed by AV-DCCA. It thus turns out that DCCA can enhance robustness and accuracy of speech recognition in noisy conditions. On the other hand, different from the first comparison, AV-DCCA has slightly less accuracy than A-DCCA. This is because of the difference of the use of audio-visual modalities. In E2E-MM, we simply introduced the visual modality in addition to the audio modality in the feature extraction, and concatenated both feature vectors as so-called feature fusion. Visual information can thus easily compensate the performance which was degraded by audio noise in ASR. In contrast, AV-DCCA simultaneously used audio and visual features projected into the same DCCA space. Since visual features have less information than clean audio features, it is considered that the contribution of visual DCCA features is limited in the AV-DCCA model. Finally, we also evaluated V-DCCA in which a model was trained using only visual DCCA features. We then found that its recognition performance is almost the same as ASR.

5.3.2 Experiment B

Fig 5 depicts recognition accuracy using the small training data set. Note that the recognition accuracy of lipreading is 65.00%. Our proposed models A-DCCA and AV-DCCA achieved better performance than ASR and E2E-MM, similar to Experiment A. On the contrary, this time AV-DCCA is superior to A-DCCA in all the SNRs. We also checked the results when using V-DCCA, and found AV-DCCA is better. Because the training set included only a few auditory clean data, the A-DCCA recognition model might suffer from the lack of reliable training data, so might V-DCCA. By jointly utilizing visual DCCA features, it is found that AV-DCCA was able to learn the model significantly, resulting the performance improvement.

6 CONCLUSION

In this paper, we proposed to utilize DCCA techniques for speech recognition, to improve its performance in noisy environments. Because DCCA tries to enhance the correlation between audio and visual modalities, we investigate how we can utilize this architecture for ASR. Compared to conventional audio-only ASR and early-fusion-based audio-visual ASR, our DCCA approaches basically achieved better recognition performance. In the first experiment, in the case we have enough training data, DCCA can be used to enhance noisy audio features resulting higher recognition accuracy. We carried out the second experiment and found that, if only a few training data are available, using not only audio DCCA but also visual DCCA features is a better strategy as data augmentation. Consequently, we found the effectiveness of applying DCCA with visual data to audio-only or audio-visual ASR.

As future works, we will test Connectionist Temporal Classification (CTC) for utterance-level speech recognition. Expansion of employing the DCCA architecture to the other tasks, not limited to audio-only or audio-visual speech recognition, is also attractive.

REFERENCES

A.Graves, Mohamed, A.-R., and G.Hinton (2013). Speech recognition with deep recurrent neural networks. In *Proc. ICASSP2013*.

A.Renduchintala, S.Ding, M.Wiesner, and S.Watanabe (2018). Multi-modal data augmentation for end-to-end asr. In *arXiv preprint arXiv:1803.10299v3*.

G.Andrew, R.Arora, J.Bilmes, and K.Livescu (2013). Deep canonical correlation analysis. In *Proc. ICML2013*, pages 1247–1255.

G.Paraskevopoulos, S.Parthasarathy, A.Khare, and S.Sundaram (2020). Multiresolution and multimodal speech recognition with transformers. In *arXiv preprint arXiv:2004.14840v1*.

H.I.Fawaz, G.Forestier, J.Weber, L.Idoumghar, and P.A.Muller (2018). Data augmentation using synthetic data for time series classification with deep residual networks. In *arXiv preprint arXiv:1810.02455*.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L., and Koishida, K. (2020). Mmtm:multimodal transfer module for cnn fusion. In *Proc. of CVPR*.

J.Thiemann and N.Ito, E. (2013). Demand: a collection of multichannel recordings of acoustic noise in diverse environments. In *Proc. ICA*, page 035081–035081.

K.Noda, Y.Yamaguchi, K.Nakadai, H.G.Okuno, and T.Ogata (2015). ”audiovisual speech recognition using deep learning. In *Applied Intelligence, 42(4)*, pages 722–737.

M.Shimonishi, S.Tamura, and S.Hayamizu (2019). Multimodal feature conversion for visual speech recognition using deep canonical correlation analysis. In *Proc. NCSP2019*.

P.Zhou, W.Yang, W.Chen, Y.Wang, and J.Jia (2019). Modality attention for end-to-end audio-visual speech recognition. In *arXiv preprint arXiv:1811.05250v2*.

R.Arora and K.Livescu (2012). Kernel cca for multi-view learning of acoustic features using articulatory measurements. In *Proc. MLSLP2012*.

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 (2010). Censrec-1-av: An audio-visual corpus for noisy bimodal speech recog- nition. In *Proc. AVSP2010*, pages 85–88.

T.DeVries and G.W.Taylor (2017). Improved regularization of convolutional neural networks with cutout. In *arXiv preprint arXiv:1708.04552*.

T.Makino, H.Liao, Y.Assael, B.Shillingford, B.Garcia, O.Braga, and O.Siohan (2019). Recurrent neural network transducer for audio-visual speech recognition. In *arXiv preprint arXiv:1911.04890v1*.

T.Tran, T.Pham, G.Carneiro, L.Palmer, and I.Reid (2017). A bayesian data augmentation approach for learning deep models. In *arXiv preprint arXiv:1710.10564*.

W.Feng, N.Guan, Y.Li, X.Zhang, and Z.Luo (2017a). Audio visual speech recognition with multimodal recurrent neural networks. In *IJCNN*.

W.Feng, N.Guan, Y.Li, X.Zhang, and Z.Luo (2017b). Audio visual speech recognition with multimodal recurrent neural networks. In *In Neural Networks (IJCNN), 2017 International Joint Conference on. IEEE*, pages 681–688.