EXTRACTING DEEP BOTTLENECK FEATURES FOR VISUAL SPEECH RECOGNITION

Chao Sui[1], *Roberto Togneri* [2], *Mohammed Bennamoun* [1]

[1]School of Computer Science and Software Engineering
[2]School of Electrical, Electronic and Computer Engineering
University of Western Australia, Australia
chao.sui@csse.uwa.edu.au, {roberto.togneri, mohammed.bennamoun}@uwa.edu.au

ABSTRACT

Motivated by the recent progresses in the use of deep learning techniques for acoustic speech recognition, we present in this paper a visual deep bottleneck feature (DBNF) learning scheme using a stacked auto-encoder combined with other techniques. Experimental results show that our proposed deep feature learning scheme yields approximately 24% relative improvement for visual speech accuracy. To the best of our knowledge, this is the first study which uses deep bottleneck feature on visual speech recognition. Our work firstly shows that the deep bottleneck visual feature is able to achieve a significant accuracy improvement on visual speech recognition.

Index Terms— Visual speech recognition, stacked denoising auto-encoder, deep bottleneck feature.

1. INTRODUCTION

Although audio-visual speech recognition has achieved significant improvements over audio-only speech recognition on both clean and noisy environments [1, 2, 3], how to encode speech related information in visual features is still a largely undeveloped area. Given the encouraging performance of deep learning techniques in acoustic speech recognition [4], in this paper, we propose a deep visual feature learning scheme that can replace existing hand-crafted visual features and boost visual speech accuracy.

Deep learning techniques were first proposed by Hinton et al. [5], who used the greedy, unsupervised, layer-wise pre-training scheme to solve the training difficulty of multiple hidden layer neural networks. Hinton et al. used the restricted Boltzmann machine (RBM) to model each layer of a deep belief network (DBN). Later works showed that a similar pre-training scheme can also be used by stacked auto-encoders [6] and convolutional neural networks (CNN) to build the deep neural network [7].

Although the speech recognition community has witnessed some great successes in the utilisation of deep learning techniques, the progress of visual speech recognition (VSR) based on deep learning is still limited. Ngiam et al. [8] first explored the possibility of applying deep networks on VSR. In their work, however, the deep auto-encoder features were used to train a support vector machine, which did not take the dynamic characteristics of speech into account. Consequently, their proposed feature learning scheme was not able to be used on practical speech recognition tasks. Huang et al. [9] trained a DBN to predict posterior probability of HMM states given the observations, which was further used for continuous speech recognition. However, the performance of their proposed visual feature learned by deep learning techniques did not show any improvements over the HMM/GMM model. Although the hand-crafted visual features still play a dominant role in VSR [3], deep learning techniques offer potential opportunities for replacing these hand-crafted features which will boost speech recognition accuracy.

In this paper, we propose an augmented deep bottleneck feature (DBNF) extraction method for visual speech recognition. Although the DBNF was extensively evaluated for the acoustic speech recognition in recent years [10, 11, 12, 13], to the best of our knowledge, this method has never been explored in visual speech recognition. In this work, a DBNF is first learned by a stacked auto-encoder and fine-tuned by a feed-forward neural network. Then, this DBNF is concatenated with the DCT feature vector, and the dimension of this concatenated feature vector is further reduced using LDA. Experimental results show that our proposed deep feature learning scheme is able to boost speech accuracy significantly.

The rest of this paper is organised as follows: Section 2 describes the proposed model for visual feature learning. The system performance is evaluated in Section 3. Finally, the paper is concluded in Section 4.

2. DEEP BOTTLENECK FEATURES

The proposed deep bottleneck visual feature extraction architecture is illustrated in Fig .1. The training process consists of three stages. The first stage is a stacked auto-encoder which is pre-trained by the video data in a layer-wise, unsupervised manner. Then, this network is further fine-tuned by adding a hidden layer and a classification layer to predict the class labels (i.e., the states of the HMMs). Finally, the deep bottle-

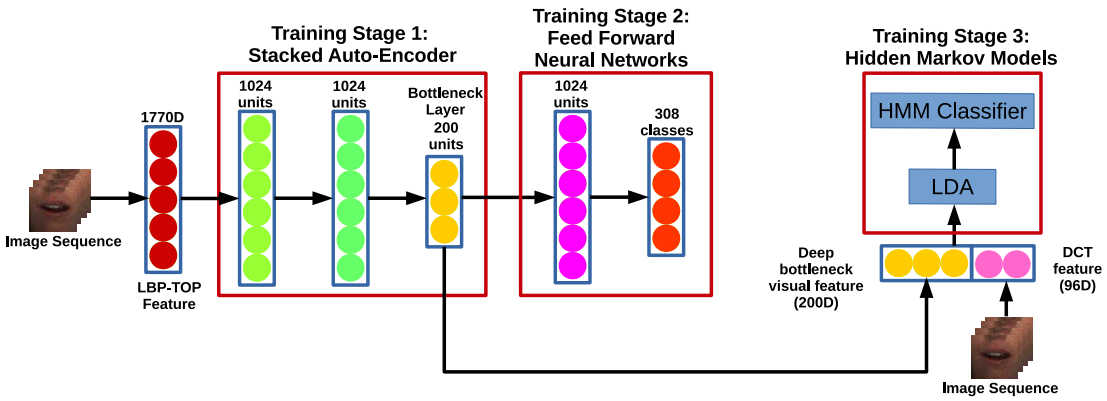

Fig. 1: Proposed augmented deep multimodal bottleneck visual feature extraction scheme.

neck feature vector is concatenated with discriminant cosine transform (DCT) feature vector, followed by a linear discriminant analysis (LDA) to decorrelate the feature and reduce the feature dimension to 20.

In terms of feature extraction, two features are extracted: DCT [14] and LBP-TOP [15]. For the DCT feature, 32 low-frequency DCT coefficients are selected in a zig-zag left to right scanning pattern, along with the 32 first and 32 second temporal derivatives to capture the dynamic information of utterances. For the LBP-TOP feature, we use a mouth region subdivision scheme introduced in [15] to extract LBP-TOP features. To be more specific, the mouth region is divided into 2×5 subregions and the 177 dimensional LBP-TOP feature vector is extracted from each of these 10 subregions to form a 1770 dimensional LBP-TOP feature vector.

Since each DCT feature element is a representation of the entire mouth region at a particular frequency, DCT is considered as a global feature representation. On the other hand, the LBP-TOP extracts local information within a small neighbourhood from both the spatial and the temporal domains. Hence, LBP-TOP is a local information representation [15]. Given the different characteristics of two appearance-based visual features in the sense of information representation, combining these two complementary information sources should be able to boost visual speech accuracy. However, compared with the 96 dimensional DCT feature, the 1770 dimensional LBP-TOP feature is not compact enough for the HMMs to perform classification. In this paper, we propose a deep feature learning based method to generate an augmented feature which embeds both global and local information into one single feature vector.

In the first stage of the stacked auto-encoder, it is a deep neural network consisting of multiple auto-encoders in which the output of each auto-encoder is wired to the input of the successive auto-encoder. For each individual layer of the stacked auto-encoder, we use a denoising auto-encoder [16] to capture the structure of the video data. The input $\mathbf{x}$ is firstly corrupted by using $\tilde{\mathbf{x}} \sim q_D(\tilde{\mathbf{x}}|\mathbf{x})$ to yield a corrupted input $\tilde{\mathbf{x}}$, where q_D is a stochastic process which randomly sets a fraction of elements of the clean input to zero.

With the corrupted input $\tilde{\mathbf{x}}$, the latent representation $\mathbf{y}$ is constructed through the encoder using the weights $\mathbf{W}$ and the bias $\mathbf{b}$ of the hidden layer and non-linear activation function σ_y:

$$\mathbf{y} = \sigma_y(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b}). \tag{1}$$

For the decoding process, the reconstruction of the input $\tilde{\mathbf{z}}$ is obtained by using Equation 1 with the transposed weight matrix $\mathbf{W}^T$ as the new weight and the bias of the visible layer $\mathbf{c}$.

The training of the denoising auto-encoder is carried out using the back-propagation algorithm to minimize the loss function $L(\mathbf{x}, \mathbf{z})$ between the clean input $\mathbf{x}$ and the reconstruction $\mathbf{z}$. For the first layer of the stacked auto-encoder, it models the LBP-TOP feature, and the mean square error is used for the loss function:

$$L(\mathbf{x}, \mathbf{z}) = \sum_{i}^{n} (\mathbf{x}_i - \mathbf{z}_i)^2, \tag{2}$$

where $i = \{1, 2, .., n\}$, and n is the number of input samples. Since the following layers of the stacked auto-encoder model the probabilities of the hidden units of the corresponding previous layers, the cross-entropy error is used as a loss function:

$$L(\mathbf{x}, \mathbf{z}) = \sum_{i}^{n} [\mathbf{x}_i \log \mathbf{z}_i + (1 - \mathbf{x}_i) \log(1 - \mathbf{z}_i)]. \tag{3}$$

The stacked auto-encoder is trained in a greedy layer-wise manner. To be specific, the first layer is first trained to minimize the error $L(\mathbf{x}, \mathbf{z})$ between the 1770-dimensional LBP features and the reconstruction of the corrupted input $\tilde{\mathbf{z}}$. Then, the corrupted activations of the first hidden units are used as the input to train the second layer. This process is repeated until the subsequent layers are pre-trained.

After the unsupervised pre-trained stage, we employ the network fine-tuning strategy proposed in [12]. More specifically, a feed-forward neural network is constructed by adding a hidden and a classification layer. In this network, the initial weights of the auto-encoder layers are obtained from the pre-training stage, and two newly added layers are initialised using random weights sampled uniformly. For the classification layer, it uses a softmax function to predict the class (i.e., the state of the HMM), and this feed-forward neural network is trained using the back-propagation algorithm.

At the third stage of the training process, the bottleneck feature is then concatenated with the DCT features to form an augmented feature (DBNF+DCT). LDA is used to decorrelate the DBNF+DCT feature vector and to further reduce the feature dimension to 20. Finally, this augmented feature is fed into an HMM recogniser.

3. EXPERIMENTS

3.1. Data Corpus

The data corpus used in our paper was collected through an Australia wide research project called AusTalk [17, 18, 19]. It is a large 3D audio-visual database of spoken Australian English, including isolated words, digit sequences, and sentences, recorded at 15 different locations in all states and territories of Australia. In the proposed work, only the digit sequence data subset is used. This set of 12 four-digit strings, which are chosen randomly to simulate the PIN recognition and telephone dialling tasks (see Table 1), is carefully designed to ensure that each digit (0-9) occurs at least once in each serial position.

Table 1: Digit sequences in the Big ASC data corpus. For the digit '0', there are two possible pronunciations: 'zero' ('z') and 'oh' ('o').

No.	Content	No.	Content	No.	Content
01	z123	02	942o	03	6785
04	123z	05	7856	06	2o94
07	23z1	08	49o2	09	8567
10	3z12	11	5678	12	0429

3.2. Experimental Setup

With the use of the method detailed in [20], the videos which capture the speakers' lip movements can be obtained. Then, the corresponding visual features can be extracted. In our experiments, we partitioned the 125 speakers into 10 non-overlapping subsets, and a 10-fold cross validation was employed. For each fold, 8 subsets of data are used for training and 2 subsets are used for testing. We run our experiments in a speaker-independent scenario; therefore the speakers in the training and test subsets do not overlap.

In order to pre-train the stacked auto-encoder, a mini-batch gradient descent with a batch size of 64 and a learning rate of 0.01 is used. A random 20% of the input elements are corrupted to zero by applying masking noise. Each layer of the stacked auto-encoder has 1024 hidden neurons, and the training of each layer is performed in 50 epochs.

After the pre-training of the stacked auto-encoder, another 1024-unit hidden layer and a classification layer are added. The whole network is then fine-tuned using a mini-batch gradient descent with a batch size of 256 and a learning rate of 0.05. Both pre-training and fine-tuning processes are carried out on GPUs and implemented by the Theano toolkit [21].

With respect to the HMM model, we use 11 word models with 30 states to model 11 digit pronunciations. Each HMM state is modelled by 9-mixture GMMs with diagonal covariance. In our experiment, the digit recognition task is treated as a connected word speech recognition problem with a simple syntax, i.e., any combination of digits and silence is allowed in any order. The HMM is implemented by the HTK toolkit [22].

3.3. Stacked Auto-Encoder Architecture

In order to confirm whether the deep feature learning architecture is necessary and can learn a better information representation than the shallow feature learning techniques, we evaluate the features that are learned by different stacked auto-encoders with various numbers of hidden layers. Meanwhile, in order to confirm that the pre-training process can benefit the visual feature learning, a stacked auto-encoder without unsupervised pre-training is also evaluated.

Table 2: Evaluation on various stacked denoising auto-encoder architectures.

Auto-Encoder Layers	Pre-training	Accuracy
1 (200)	Yes	43.2%
2 (1024-200)	Yes	55.7%
3 (1024-1024-200)	**Yes**	**57.3%**
3 (1024-1024-200)	No	49.9%
4 (1024-1024-1024-200)	Yes	57.1%

Table 2 reports the visual speech accuracy using the features learned by various stacked auto-encoder architectures. From this table, one can observe that with an increase in the number of hidden layers, a better feature representation can be obtained. Meanwhile, one can also note that the use of pre-training results in a better accuracy. However, the table also shows that, with 3 hidden layers, increasing the hidden layers is not able to further boost speech accuracy. Similar results were also found in the acoustic speech recognition tasks [12]. Specifically, with a sufficiently large number of auto-encoder layers, increasing the number of layers cannot further boost the speech recognition accuracy. A possible explanation is

Table 3: Visual speech recognition performance comparison between our proposed DBNF and other methods.

Feature	Reduction	Dimension	Accuracy
DCT	MMI	60	52.3%
DCT	mRMR	60	52.2%
DCT	CMI	80	51.1%
DCT	LDA	20	54.7%
LBP-TOP	MMI	190	52.5%
LBP-TOP	mRMR	190	53.0%
LBP-TOP	CMI	310	37.0%
DBNF	**None**	**200**	**57.3**%
DBNF	**LDA**	**40**	**63.3**%
Augmented DBNF	**LDA**	**20**	**67.8**%

that when the auto-encoder is large enough, adding new layers cannot increase the representative ability of the network. Moreover, adding new layers requires a larger amount of data to ensure the auto-encoder is sufficiently trained.

3.4. Performance of the Augmented Bottleneck Feature

Unlike the standard bottleneck feature learning process, the learned bottleneck feature is concatenated with the DCT feature, and LDA is further used to decorrelate the feature vector and to reduce the feature vector dimensionality. Hence, the superiority of this feature extraction scheme needs to be evaluated.

As illustrated in Table 3, with the use of LDA, the accuracy of the DBNF increases from 57.3% (200 dimensions) to 63.3% (40 dimensions), which shows that LDA is able to decorrelate the feature learned by the stacked auto-encoder and reduce the feature dimension. Meanwhile, our proposed method yields an accuracy of 67.8 % by concatenating the DCT feature with the DBNF. It shows that our proposed augmented DBNF is able to produce an even higher accuracy because it embeds both local and global information into one single feature vector.

In order to demonstrate the superiority of our proposed augmented DBNF, we list some other popular appearance-based visual features in Table 3. Particularly, we compare our proposed DBNF and augmented DBNF with two features (DCT [14] and LBP-TOP [15]) and two feature reduction techniques, i.e., LDA [1] and mutual information feature selector (MMI, mRMR, CMI) [14]. As shown in Table 3, the visual speech accuracy of our proposed augmented DBNF, which takes two complementary information representation methods (i.e., local and global information) into account, outperforms all the listed visual feature types and feature dimension reduction schemes.

More specifically, as shown in Table 3, besides the proposed deep learning techniques in this work, DCT with LDA (DCT+LDA) yields the highest accuracy (54.7%). In our study, we also found that LDA failed to obtain a proper transformation on the raw 1770-dimensional LBP-TOP feature vectors, because modelling such a dimensional feature using LDA requires that there are at least 1770 training samples for each of the 308 classes (HMM states). Although the data corpus we used is a relatively larger audio-visual connected digit speech database, the amount of data is still not large enough to perform the LDA reduction on the LBP-TOP features. Compared with the mutual information feature selectors (MMI, mRMR and CMI), using the proposed stacked auto-encoders to learn features achieves a relative improvement of 8% (57.3%), because the deep learning techniques are able to make full use of the information in the LBP-TOP feature, while the mutual information selectors only select several relatively informative components from the original LBP-TOP.

Since LDA is able to decorrelate the feature components and reduce the feature dimension, we use LDA to further optimize the DBNF. After the optimization, the visual speech accuracy further increases to 63.3%. The reason for the significant accuracy increase are two fold: 1) Compared with the 200-dimensional DBNF, the feature dimension is dramatically reduced to 40 to avoid the curse of dimensionality. 2) Since the units of the stacked auto-encoder are fully connected between the units in adjacent layers, the components of the DBNF are correlated. Employing the LDA is able to decorrelate the components in the DBNF. In this work, we also proposed an augmented DBNF, which is able to embed both local (LBP-TOP) and global (DCT) information into a compact feature vector, and our proposed augmented DBNF yields 24% relative improvement, compared with DCT+LDA.

4. CONCLUSION

In this paper, we propose an augmented DBNF for visual speech recognition. This augmented DBNF is first learned with a stacked denoising auto-encoder, followed by a fine-tune process using a feed forward neural network. The DBNF is then augmented by concatenating the DCT feature vector, and LDA is applied to decorrelate the feature and reduce the feature dimension. Experimental results show that our proposed augmented DBNF significantly boosts speech accuracy. Unlike the recently proposed works which solve the lipreading problems as a classification problem [23, 24, 25, 26], we tackled this problem similar to a speaker-independent acoustic speech recognition task, which needs to capture the temporal dynamic of the data (e.g. by using HMMs). To the best of our knowledge, this is the first work which explores the use of the deep bottleneck feature on visual speech recognition, and firstly show that the deep learned visual features can achieve a significant improvement than the hand-crafted features.

5. REFERENCES

[1] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[2] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, pp. 23, 2004.

[3] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong, and Matti Pietikäinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, 2014.

[4] Li Deng and Dong Yu, *Foundations and Trends in Signal Processing: DEEP LEARNING — Methods and Applications*, Microsoft Research, June 2014.

[5] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[6] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, pp. 153, 2007.

[7] M Ranzato, Fu Jie Huang, Y-L Boureau, and Yann LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[8] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 689–696.

[9] Jing Huang and Brian Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7596–7599.

[10] Dong Yu and Michael L Seltzer, "Improved bottleneck features using pretrained deep neural networks.," in *INTERSPEECH*, 2011, pp. 237–240.

[11] Tara N Sainath, Brian Kingsbury, and Bhuvana Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4153–4156.

[12] Jonas Gehring, Yajie Miao, Florian Metze, and Alex Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3377–3381.

[13] Bing Jiang, Yan Song, Si Wei, Jun-Hua Liu, Ian Vince McLoughlin, and Li-Rong Dai, "Deep bottleneck features for spoken language identification," *PloS one*, vol. 9, no. 7, 2014.

[14] Mihai Gurban and J Thiran, "Information theoretic feature extraction for audio-visual speech recognition," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4765–4776, 2009.

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

[16] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[17] Denis Burnham, Eliathamby Ambikairajah, Joanne Arciuli, Mohammed Bennamoun, Catherine T Best, Steven Bird, AB Butcher, C Cassidy, Girija Chetty, Felicity M Cox, et al., "A blueprint for a comprehensive Australian English auditory-visual speech corpus," in *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus, ed. Michael Haugh et al*, 2009, pp. 96–107.

[18] Michael Wagner, Dat Tran, Roberto Togneri, Phil Rose, David Powers, Mark Onslow, Debbie Loakes, Trent Lewis, Takaaki Kuratate, Yuko Kinoshita, et al., "The big Australian speech corpus (the big ASC)," in *Proceedings of 13th Australasian International Conference on Speech Science and Technology*, 2010, pp. 166–170.

[19] Denis Burnham, Dominique Estival, Steven Fazio, Jette Viethen, Felicity Cox, Robert Dale, Steve Cassidy, Julien Epps, Roberto Togneri, Michael Wagner, et al., "Building an audio-visual corpus of Australian English: Large corpus collection with an economical portable and replicable black box," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[20] Chao Sui, Serajul Haque, Roberto Togneri, and Mohammed Bennamoun, "A 3D audio-visual corpus for speech recognition," in *Proceedings of Australasian International Conference on Speech Science and Technology*, 2012.

[21] James Bergstra, Olivier Breuleux, and Frédéric Bastien, "Theano: a cpu and gpu math expression compiler," in *Proceedings of the Python for scientific computing conference (SciPy)*, 2010.

[22] Steve J Young, Gunnar Evermann, MJF Gales, D Kershaw, G Moore, JJ Odell, DG Ollason, D Povey, V Valtchev, and PC Woodland, "The HTK book version 3.4," 2006.

[23] Amr Bakry and Ahmed Elgammal, "Mkpls: Manifold kernel partial least squares for lipreading and speaker identification," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 684–691.

[24] Yuru Pei, Tae-Kyun Kim, and Hongbin Zha, "Unsupervised random forest manifold alignment for lipreading," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 129–136.

[25] Ziheng Zhou, Xiaopeng Hong, Guoying Zhao, and Matti Pietikainen, "A compact representation of visual speech data using latent variables," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 1, Jan 2014.

[26] Jingyong Su, Anuj Srivastava, Fillipe DM de Souza, and Sudeep Sarkar, "Rate-invariant analysis of trajectories on riemannian manifolds with application in visual speech recognition," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014.