This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Listening With Your Eyes: Towards a Practical Visual Speech Recognition System Using Deep Boltzmann Machines

Chao Sui, Mohammed Bennamoun
School of Computer Science and Software Engineering
University of Western Australia, Perth, Australia
chao.sui@csse.uwa.edu.au, mohammed.bennamoun@uwa.edu.au

Roberto Togneri
School of Electrical, Electronic and Computer Engineering
University of Western Australia, Perth, Australia
roberto.togneri@uwa.edu.au

Abstract

This paper presents a novel feature learning method for visual speech recognition using Deep Boltzmann Machines (DBM). Unlike all existing visual feature extraction techniques which solely extract features from video sequences, our method is able to explore both acoustic information and visual information to learn a better visual feature representation. During the test stage, instead of using both audio and visual signals, only the videos are used to generate the missing audio features, and both the given visual and audio features are used to produce a joint representation. We carried out experiments on a new large scale audio-visual corpus, and experimental results show that our proposed techniques outperform the performance of hand-crafted features and previously learned features and can be adopted by other deep learning systems.

1. Introduction

Continuous efforts have been made towards the development of Automatic Speech Recognition (ASR) systems in the recent years, and numerous ASR systems (e.g. Apple Siri and Microsoft Cortana) have come into use in our daily life. Although ASR research has made remarkable progress, practical ASR systems are still prone to environmental noises. A possible solution to overcome the recognition degradation in the presence of acoustic noises is to take advantage of the visual stream which is able to provide complementary information to the acoustic channel. Despite the promising application prospects of Audio-Visual Speech Recognition (AVSR), the problem on how to extract visual features from videos still remains a difficult one. In order to improve visual feature representation techniques, Visual Speech Recognition (VSR), also known as lipreading, have emerged as an attractive research area in the recent years [31].

Figure 1. Possible application scenarios of our proposed framework. In an noisy environment, visual features are a promising solution for automatic speech recognition.

features to boost speech accuracy, another promising aspect of VSR is in its wider potential real-life applications compared to acoustic based ASR. As shown in Fig. 1, in many practical applications, ASR systems are exposed to noisy environments, and the acoustic signals are almost unusable for speech recognition. On the other hand, with the availability of front and rear cameras on most mobile devices, users can easily record facial movements to improve speech accuracy. In these extremely noisy environments, the visual information becomes basically the only source that ASR systems can use for speech recognition.

Although lipreading techniques provide an effective potential solution to overcome environmental noises for ASR systems, there are still several challenges in this area. Unlike the well-established audio features, such as the Mel Frequency Cepstral Coefficients (MFCC), how to encode the speech-related visual information into a compact feature vector is still a difficult problem, because the lip movements are not easily distinguishable compared to audio signals between different utterances. Another challenge is that the fusion of both audio and visual signals dramatically degrades the speech recognition performance in the presence of the noisy acoustic signals.

Given the aforementioned visual speech recognition challenges, this paper provides a new perspective to solve both of these challenges. For the first challenge, since the audio features perform much better than the visual features, we use both audio and visual information to learn a more pertinent feature representation for speech recognition. Moreover, the trained feature representation model is also capable of inferring the audio features when visual information is available. Hence, during the test stage, the audio signals which may be severely corrupted by the environmental noises are not required. Instead, the visual feature is used to reconstruct the audio information, and both the given visual and inferred audio information are used to yield a joint feature representation. Therefore, the second aforementioned challenge can be solved.

The rest of this paper is arranged as follows: Section 2 introduces some related works, and based on the review of relevant recent works, the contributions of this paper are also given in this section. The feature learning scheme is presented in Section 3. We extensively evaluate the performances of different visual features in Section 4. Finally, we summarize our paper in Section 5.

2. Related Works and Our Contributions

Compared with the well established audio features e.g. MFCC, there is no universally accepted visual feature to represent lipreading relevant information [31]. In this section we first review the recent visual feature extraction works. We then highlight the key contributions of our work to this area.

Generally speaking, visual features can be divided into four categories: appearance-based features and shape-based features [17]. For appearance-based features, image transformation techniques are performed on raw image pixels to extract visual features, while the parameters of the lip shape models are used to extract the shape-based features. Although the shape-based visual features are able to explicitly capture the shape variations of the lips, an extremely large number of lip landmarks need to be laboriously labelled, which is infeasible for large-scale speech recognition tasks. On the other hand, appearance-based visual features are computationally efficient and do not require any training process. Hence, appearance-based features have been widely adopted in the recent years [31].

In terms of appearance-based visual features, Zhao et al. [29] introduced a Local Binary Pattern (LBP) based spatial-temporal visual feature, called LBP-TOP. This feature produced an impressive performance over other existing feature extraction techniques on various lipreading tasks. Despite the promising performance of LBP-TOP features, the dimensionality of the raw LBP-TOP feature is very large, which makes the system succumb to the curse of dimensionality. Hence, a number of other works have been presented to encode the visual information of the LBP-TOP features by more informative representations [32, 13, 1, 15, 22, 30]. However, these works focused only on isolated words and phrase recognition. They did not consider connected words or continuous speech recognition, which is highly in demand by modern speech recognition systems using Hidden Markov Models (HMMs) [2]. Given the rich speech relevant visual information embedded in LBP-TOP visual features, this paper presents a novel feature learning technique which can explore the speech relevant information from the raw LBP-TOP features.

Motivated by the great success achieved by deep learning techniques in the area of acoustic speech recognition [3], this paper introduces a new visual feature learning technique to improve lipreading accuracy. In this paper, we use Deep Boltzmann Machines (DBM) [18] to learn the visual features. Encouraging pioneering works, which employed deep learning techniques for visual speech recognition, have been carried out by Ngiam et al. [12] and Huang et al. [8]. However, in [12], the visual features trained by the deep Auto-Encoder (AE) were fed to a Support Vector Machine (SVM), which limited the work to mainly isolated word recognition. Huang et al. [8] trained a Deep Belief Network (DBN) to predict the posterior probability of HMM states given the observations, which can be used for continuous speech recognition. However, the performance of their proposed visual feature learned by deep learning techniques did not show marginal improvements over the benchmark HMM/GMM model.

Meanwhile, Ngiam et al. [12] proposed a cross modality learning framework that used both audio and visual information to train a shared representation. However, this framework failed to yield a better accuracy than the visual-only learning framework. This cross modality learning framework provides, however, a new perspective to overcome the low lipreading accuracy problem. More specifically, although practical ASR systems are usually exposed to acoustic noisy environments, it is always easy to collect both clean audio and visual data in controlled lab environments. This means that we can train a feature learning model that uses both clean audio and visual signals to learn a better shared representation. When this well trained system is used under noisy environment, instead of relying on the noisy acoustic signals, only the captured video signal are used to generate the joint feature representation for visual speech recognition. The feature learning techniques used in [12, 8] are not able to generate an adequate shared representation when one modality (i.e. audio) is missing.

Fortunately, Salakhutdinov and Hinton [18] proposed a Deep Boltzmann Machine (DBM) which is a deep Restricted Boltzmann Machine (RBM) based model with the ability to infer a missing modality. Unlike DBN [7] which is a directed model based on RBM, DBM is an undirected graphical model with bipartite connections within adjacent hidden layers. The undirected structure of the DBM allows this model to infer a missing modality by a Gibbs sampler [18]. Srivastava and Salakhutdinov firstly used DBM on multimedia images and text tags multimodal classification [20], and later demonstrated the superior performance of DBM in the case of audio-visual speech recognition [21]. However, how to infer the missing audio from the video was not considered in their paper.

In this paper we propose a novel formulation of the multimodal DBM to the audio-visual connected word speech recognition task and propose the following key contributions:

- Unlike previous works that only extract visual features from video data, we propose a novel framework that uses both the audio and visual signals to enrich the visual feature representation.
- Although both audio and visual features are required in the training, we only use the visual features for the testing since our feature learning framework is capable of inferring the missing (i.e. degraded) audio modality.

Hence our proposed framework provides a promising solution for practical automatic speech recognition systems. It is deployed in very noisy environments and exploits the more reliable visual modality instead of audio signals.

3. Proposed Feature Learning Scheme

The block diagram of proposed system is shown in Fig. 2. The visual feature learned by the Deep Boltzmann Machine (DBM) is concatenated with Discrete Cosine Transform (DCT) feature vector, followed by a Linear Discriminant Analysis (LDA) to decorrelate the feature and reduce the feature dimension. Then, the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) is used as a classifier for visual speech recognition. In the following section, the DBM model is first introduced.

3.1. Multimodal Deep Boltzmann Machine

The DBM consists of a series of Restricted Boltzmann Machine (RBM) stacked on top of each others. The energy of the joint configuration of the visible units $\boldsymbol{v}$ and hidden units $\boldsymbol{h}$ can be formulated as:

$$E(\boldsymbol{v},\boldsymbol{h};\theta) = -\boldsymbol{v}^\top \boldsymbol{W}^{(1)}\boldsymbol{h}^{(1)} - \sum_{i=2}^{n} \boldsymbol{h}^{(i-1)\top}\boldsymbol{W}^{(i-1)}\boldsymbol{h}^{(i)}, \tag{1}$$

where $\theta = \{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, ..., \boldsymbol{W}^{(n-1)}\}$ is the model parameter, which is the set of weights between the different layers. The joint distribution of the model can be formulated as:

$$\begin{aligned} P(\boldsymbol{v};\theta) &= \sum_{\boldsymbol{h}^{(1)},...,\boldsymbol{h}^{(n)}} P(\boldsymbol{v},\boldsymbol{h}^{(1)},...,\boldsymbol{h}^{(n)};\theta) \\ &= \frac{1}{Z(\theta)} \sum_{\boldsymbol{h}^{(1)},...,\boldsymbol{h}^{(n)}} exp(-E(\boldsymbol{v},\boldsymbol{h}^{(1)},...,\boldsymbol{h}^{(n)};\theta)), \end{aligned} \tag{2}$$

where $Z(\theta)$ is the partition function. The training process of DBM can be divided into two steps: pre-training and fine-tuning, which will be introduced in the following.

3.2. DBM Pre-training

The pre-training of the DBM is carried out by training RBMs in a greedy layer-wise manner. Since the inputs of the DBM are real-valued and all the hidden units are binary, the RBMs between the input layer and the hidden layer are Gaussian RBMs, while the RBMs between the adjacent hidden layers are binary RBMs.

The DBM has two real-valued input streams: the visual input $\boldsymbol{v}_v$ and the audio input $\boldsymbol{v}_a$, and a sequence of binary-valued hidden layers. For the D-dimensional input $\boldsymbol{v}_i \in \{\boldsymbol{v}_v, \boldsymbol{v}_a\}$ of each stream, the energy of the D-dimensional input $\boldsymbol{v}$ and the first layer $\boldsymbol{h}^{(1)}$, which consists of F hidden units, can be modelled as follows:

$$\begin{aligned} E(\boldsymbol{v},\boldsymbol{h}^1;\theta) = \sum_{i=1}^{D} \frac{(v_i - b_i)^2}{\sigma_i^2} - \sum_{i=1}^{D}\sum_{j=1}^{F^{(1)}} \frac{v_i}{\sigma_i} W_{ij} h_j^{(1)} \\ - \sum_{j=1}^{F^{(1)}} a_j h_j^{(1)}, \end{aligned} \tag{3}$$

where $\theta = (\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$ are the model parameters, $\boldsymbol{W}$ is the weight between two adjacent layers, $\boldsymbol{a}$ is the bias of the hidden layer, $\boldsymbol{b}$ is the bias of the visible layer, and σ is the standard deviation of the input.

The energy between the k hidden layer and the $k+1$ hidden layer is defined by Eq. 4. This process is continued until all RBMs layers are pre-trained using Contrastive Divergence (CD) [7]. Once both the audio and visual streams are pre-trained separately, an additional hidden layer is added on top of audio and visual streams.

$$\begin{aligned} E(\boldsymbol{h}^{(k)},\boldsymbol{h}^{(k+1)};\theta) = -\sum_{i=1}^{F^{(k)}}\sum_{j=1}^{F^{(k+1)}} h_i^{(k)} W_{ij} h_j^{(k+1)} \\ - \sum_{i=1}^{F^{(k)}} h_i^{(k)} b_i - \sum_{j=1}^{F^{(k+1)}} a_j h_j^{(k+1)} \end{aligned} \tag{4}$$

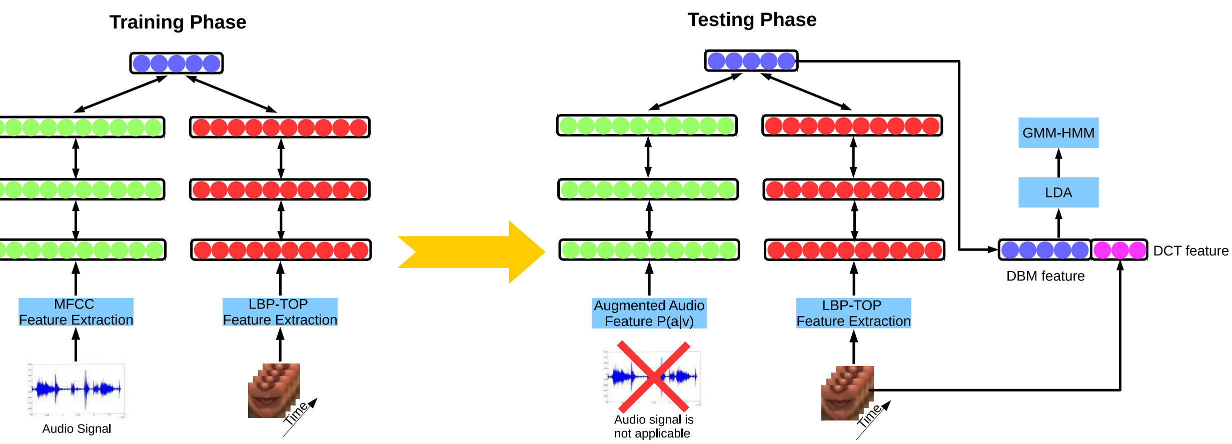


Figure 2. Block diagram of our proposed system. The left side of the figure shows the training phase.The visual feature is learned from both the audio and video stream using a multimodal DBM. The right side of the figure shows the testing phase, where the audio signal is not used. In the testing phase, the audio is generated by clamping the video input and sampling the audio input from the conditional distribution.

3.3. DBM Fine-tuning

Once the units in each layer are pre-trained, the joint distribution of the multimodal DBM, as shown in Fig. 2, is formulated by applying the visual (1st) and visual-hidden (2nd) terms in Eq. 3, and the hidden-hidden(1st) term from Eq. 4, and using Eq. 2 to yield the joint distribution over the audio and visual inputs:

$$P(\boldsymbol{v};\theta) = \frac{1}{Z(\theta)} \sum_{\boldsymbol{h}^{(2)},\boldsymbol{h}^{(3)}} exp(\sum_{k\in(a,v)} (-\sum_{i=1}^{D_k} \frac{(v_{ki}-b_{ki})^2}{\sigma_{ki}^2} + \sum_{i=1}^{D_k}\sum_{j=1}^{F_k^{(1)}} \frac{v_{ki}}{\sigma_i} W_{kij} h_{kj}^{(1)} + \sum_{i=1}^{F_k^{(1)}}\sum_{j=1}^{F_k^{(2)}} h_{ki}^{(1)} W_{kij} h_{kj}^{(2)} + \sum_{i=1}^{F_k^{(2)}}\sum_{j=1}^{F_k^{(3)}} h_{ki}^{(2)} W_{kij} h_{kj}^{(3)})), \quad (5)$$

where $k \in \{a, v\}$ represents the audio(a) and audio(v) streams. The parameters of the model are fine-tuned by approximating the gradient of the log-likelihood of the probability that the model assigns to the visible vectors $\boldsymbol{v}$, i.e. $\mathcal{L}(P(\boldsymbol{v};\theta))$, with respect to the model parameters θ.

In [7], this process can be formulated as:

$$\frac{\partial \mathcal{L}(P(\boldsymbol{v};\theta))}{\partial\theta} = \alpha(E_{P_{data}}[\boldsymbol{v}\boldsymbol{h}^\top] - E_{P_{model}}[\boldsymbol{v}\boldsymbol{h}^\top]), \quad (6)$$

where α is the learning rate. $E_{P_{data}}[.]$ represents the data-dependent expectation, which is the expectation of the $P(\boldsymbol{v};\theta)$ with respect to the training data set. $E_{P_{data}}[.]$ represents the model expectation, which is the expectation of the $P(\boldsymbol{v};\theta)$ defined by the model (Eq. 5).

The model parameters approximation process can be divided into two separate procedures. For the data-dependent expectation estimation, the mean-field inference is used, followed by a Markov Chain Monte Carlo (MCMC) based stochastic approximation procedure to approximate the model-dependent expectation. Further details of the training process can be found in [18].

3.4. Generating Missing Audio Modality

One highlight of this paper is the introduction of a new perspective for visual feature extraction, wherein the visual feature is learned from both the visual and audio modalities. This technique provides a more practical solution for many speech recognition scenarios where the audio signals are almost unusable because of the environmental noises. However, in order to make this system feasible, the missing audio signals need to be generated by the trained DBM model during the classification. Fortunately, since the DBM is an undirected generative model, audio signals can be inferred from the visual modality. Then, the reconstructed audio feature can be used as an augmented input to perform visual speech recognition.

More specifically, given the observed visual features, the audio feature is inferred by clamping the visual feature at the input units, and applying a standard alternating Gibbs sampler [18] to sample the hidden units from the conditional distribution using the following equations:

$$P(h_j^k = 1|\boldsymbol{h}^{k-1}, \boldsymbol{h}^{k+1}) = \sigma(\sum_i W_{ij}^k h_i^{k-1} + \sum_m W_{jm}^{k+1} h_m^{k+1}), \quad (7)$$

$$P(h_m^n = 1|\boldsymbol{h}^{n-1}) = \sigma(\sum_j W_{jm}^n h_j^{n-1}), \quad (8)$$

$$P(v_i = 1|\boldsymbol{h}^1) = \sigma(\sum_i W_{ij}^i h_j^1), \quad (9)$$

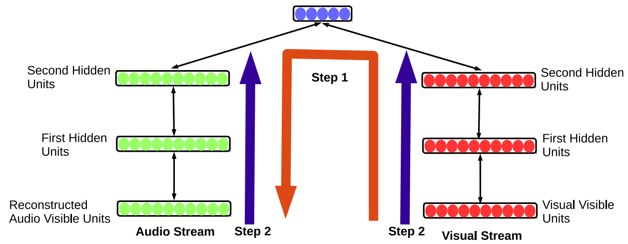


Figure 3. The generation of the missing audio signals can be divided into two steps: 1. Infer the audio signal from the given visual features. 2. Generate a joint representation using both the reconstructed audio and the given visual features.

where i, j and m are the indices of the units in the corresponding layers.

Once the audio feature is reconstructed, both the generated audio and the given visual features are used together to generate a joint representation for speech recognition. This process is illustrated in Fig. 3, and the details of this process are also shown in Algorithm 1.

Algorithm 1 Porcess of generating the missing audio feature

Clamp the observed visual feature $\boldsymbol{v}_v$ at the input.
for Each hidden layer k in visual stream. **do**
  Gibbs Sample the hidden layer state in a bottom-up manner, and estimate $P(h_i^k = 1|\boldsymbol{h}^{k-1}, \boldsymbol{h}^{k+1})$ using Eq. 7.
end for
Gibbs sample the joint layer state , and estimate $P(h_m^n = 1|\boldsymbol{h}^{n-1})$ using Eq. 8.
for Each hidden layer k in audio stream **do**
  Gibbs Sample the hidden layer state in a top-down manner, and estimate $P(h_j^k = 1|\boldsymbol{h}^{k-1}, \boldsymbol{h}^{k+1})$ using Eq. 7.
end for
Infer the missing audio feature using Eq. 9.
Gibbs Sample the joint representation in a bottom-up manner by feeding both the reconstructed audio and observed visual features into the network.

3.5. Deterministic Fine-Tuning

Once the DBM model is fully trained, its weights are used to initialise a deterministic multilayer neural network as in [18]. For each input vector $\mathbf{v}_i \in \{\mathbf{v}_v, \mathbf{v}_a\}$, the approximate posterior distribution $Q(\mathbf{h}_i|\mathbf{v}_i)$ is estimated by the mean-field inference. The marginals of $Q(\mathbf{h}_i|\mathbf{v}_i)$ are then used with the input vector $\mathbf{v}_i$ to form an augmented input for the deterministic multilayer neural network, as shown in Fig. 4. The standard backpropagation is used to discriminatingly fine-tune the model.

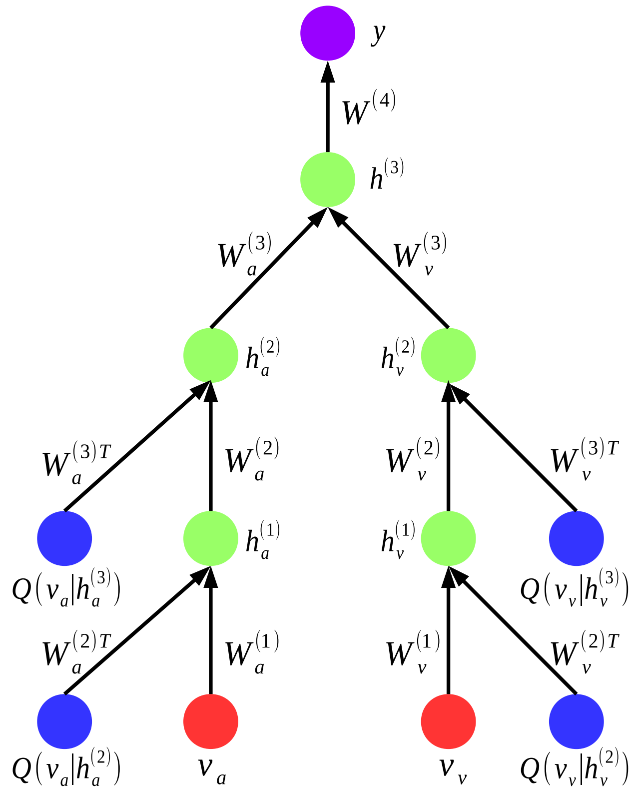


Figure 4. Discriminative fine-tuning of our proposed DBM model.

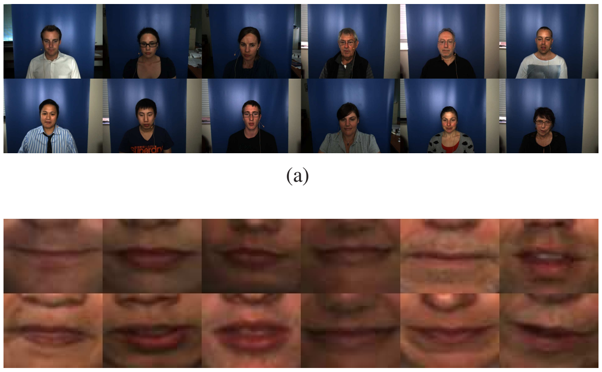
3.6. Augmented Visual Feature Generation

In the last step of our proposed method, the visual feature learned from the DBM model is then concatenated with the DCT feature to form an augmented visual feature (as shown in Fig. 2). The LDA is used to decorrelate the augmented visual feature vector and to further reduce the feature dimension. Similar feature augmentation techniques have already been used in both acoustic speech recognition [28] and visual speech recognition [23]. Finally, this augmented visual feature is fed into an HMM recogniser.

4. Experiments

4.1. Data Corpus and Experimental Setup

Many high quality recent papers mainly focused on the task of isolated word/letter recognition or phrase classification [32, 13, 12, 1, 15, 22, 30, 21] . In contrast, we propose a more practical lipreading system that can perform connected speech recognition. In this case, the popular benchmark corpora, such as AVLetters [9], CUAVE [14] and OuluVS [29] or the combination of these corpora (as used in [12, 21]) are not fully useful because they are limited in both speaker number and speech content. In addition, some large-scale data corpora such as AVTIMIT AVTIMIT [6], IBMIH [8] are not publicly accessible. Al-



(a)

(b)

Figure 5. Examples in the AusTalk Corpus. Fig. 5a: Original recordings in the corpus. Fig. 5b: Corresponding mouth ROI examples extracted from the original examples in Fig. 5a.

though XM2VTSDB [10] (which is publicly available) has 200 speakers, the speech is limited to simple sequences of isolated word and digit utterances. In order to evaluate our VSR system on the more difficult connected speech recognition problem, a new large-scale audio-visual data corpus was established and used in our paper.

The data corpus used in our paper was collected through an Australia wide research project called AusTalk [26]. It is a large-scale 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. In the digit data subset, 12 four-digit strings are provided for people to read. This set of digit strings, which are organised in a random manner without any unnatural pause to simulate the PIN recognition and telephone dialling tasks. Moreover, the digit selection was carefully designed to ensure that each digit (0-9) occurs at least once in each serial position (see Table 1).

Table 1. Digit sequences in the AusTalk 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

Some recording examples can be seen in Fig. 5a. To generate the required visual information, the mouth ROIs, as illustrated in Fig. 5b, are cropped from the original videos using the Harr features and Adaboost [25].

In order to increase the statistical significance of our results, we split all the 125 speakers' digit session recording data into three groups, i.e., the training set, the validation set, and the test set. The speakers in the different groups are not overlapped. A three-fold cross validation is then used, and the average word accuracy of the three runs are reported.

4.2. Audio and Visual Features

In terms of the audio feature, a 13 Mel Frequency Cepstral Coefficients (MFCC) with Cepstral Mean Normalisation (CMN) was extracted, and the zero-th coefficient appended. Then, each 13-dimensional MFCCs were stacked across 11 consecutive frames which results in a total of 143 coefficients for each frame, as in [4].

The visual feature used in this study is the LBP extracted from Three Orthogonal Planes or LBP-TOP [29]. Unlike the basic LBP for static images, LBP-TOP extends feature extraction to the spatial-temporal domain, which makes LBP an effective dynamic texture descriptor. More specifically, given a mouth ROI frame at a particular time, a 59-bin histogram is generated to accumulate the presence of different uniform binary patterns across the spatial and temporal planes. Then, a 177-dimensional feature vector is generated by concatenating the three histograms to represent both the lip appearance and its motion. In order to get a promising speech accuracy, the mouth region is further divided into 2×5 subregions, as elaborated in [29]. Since the visual input units of our multimodal DBM are the LBP-TOP features from 10 mouth subregions, the number of the visual input units is 1770. Compared to the direct use of the pixels of raw images at the input of the network, the LBP-TOP feature embeds both the spatial and temporal information in a single feature vector, which enables the DBM to find a feature representation that captures both appearance information and lip movements.

4.3. Learning Model Architecture

The visual stream of the multimodal DBM consists of a Gaussian RBM with 1770 visible units followed by 2 hidden layers with 2048 and 1024 units, respectively. The audio stream consists of a Gaussian RBM with 143 visible units followed by 2 hidden layers with 256 and 128 units, respectively. On the top of the visual and audio streams, there is a joint representation layer with 256 units. In order to compare the proposed multimodal DBM with the unimodal counterpart, the unimodal DBM is also set up with three hidden layers of 2048, 1024, 256 units respectively. The deep learning techniques used in this paper were implemented using DeepNet and CudaMat [11], and the learning process was implemented by exploiting the parallel pipeline architecture of NVIDIA GPU on PC desktop workstations. Since training the DBM only requires unlabelled data, we combined the data from both AusTalk and AVLetters for the feature learning.

As for the classifiers of the AusTalk data, the HMMs were implemented using the HTK software [27]. With the application of the HTK, we implemented 11 HMM word models with 30 states to model 11 digit pronunciations. Each HMM state was modelled by a 9-mixture GMM with a diagonal covariance. In our experiments, 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 to simulate the real speech driven tasks, such as telephone number dialing and telephone banking tasks. In order to obtain the class label of each frame for the discriminative learning, a forced alignment was carried out on all utterances to obtain the word boundary positions.

4.4. Performance Evaluation

In the following section, we evaluate the performance of our proposed method, and compare our model with other popular feature extraction and learning methods. We first demonstrate the missing modality inference ability of our model.

As explained in Section 3, one highlight of our proposed method is the ability of our model to use the inferred audio features to boost the visual speech accuracy. Hence, we used the visual features together with three different audio features, i.e., clean audio, inferred audio and zero paddings, to generate the DBM learned feature. We then used this feature to train the HMM classifier, and listed our recognition results in Table 2.

Table 2. Connected digit recognition performance with multimodal inputs.

Model	Audio Input	Accuracy
Deep Boltzmann Machine	Zero padding	34.2%
	Clean audio	79.8%
	Inferred audio	59.9%

Table 2 shows that the use of clean audio features is able to generate a very promising result (79.8%). On the other hand, using an all zeros vector to replace the audio features fails to produce a good accuracy (34.2%), because the DBM model was trained with both visual and audio features and the use of fake audio signals (with all zeros) makes the model fail to work properly. Hence, in order to generate a useful joint feature representation from the DBM model, both visual and audio signals need to be provided.

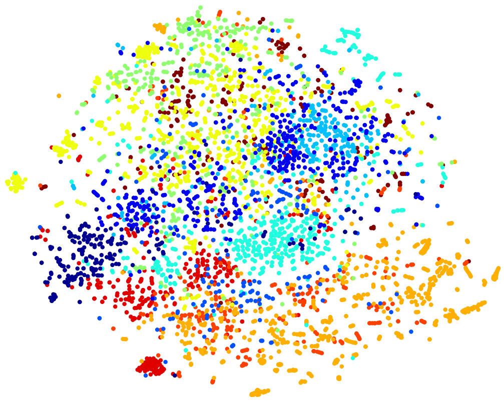
Once the audio features are inferred from the visual features, we used them with the visual features to generate joint representations. Table 2 shows that although the accuracy produced by the inferred audio is not as good as its clean audio counterpart (79.8% vs 59.9%), the accuracy is still very promising compared to other VSR methods, because the inferred audio is able to provide additional information that is helpful to speech recognition. The ability of our method to infer the missing audio provides a new potential solution for AVSR systems. More specifically, previous AVSR mainly focus on the dynamic assignment of different weights to the audio and video streams according to the noise level of the audio signals. These AVSR systems usually fail to achieve a satisfactory accuracy because of their inaccurate estimation of the noise level. In contrast, our proposed work reconstructs the clean audio from the multimodal DBM, rather than using the audio signal with unknown noise levels.

As shown in Fig. 2, once the feature is obtained from the DBM model, it is concatenated with the corresponding DCT feature to generate an augmented visual feature. LDA is then used to reduce the feature dimension. Similar frameworks were also used in both acoustic speech recognition [28] and visual speech recognition [23]. In order to show the necessity of the augmented DBM feature, a set of experiments were carried out, and the results are listed in Table 3. From this table, one can note that using LDA to reduce the dimensionality of the DBM learned features before feeding them into the HMM is able to produce a higher accuracy (64.4%) compared to using the DBM learned feature directly on the HMM (59.9%). Furthermore, combining the DBM learned feature with the DCT feature produces the highest accuracy (69.1%), because our proposed augmented visual feature combines different types of visual information. More specifically, the DBM learns the feature from the LBP-TOP feature which is a local information representation, while the DCT is a global feature representation [29]. Hence, our proposed feature learning method is able to represent the speech information at both the local and global levels, thereby producing a superior accuracy.

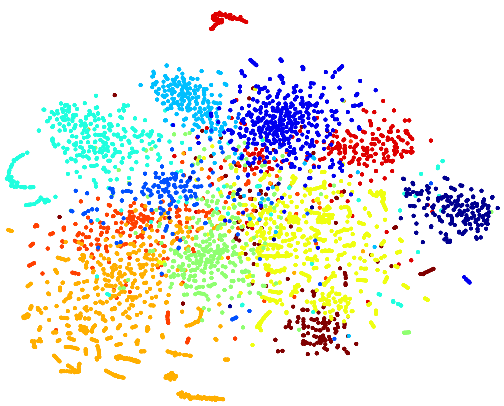
Table 3. Performance comparison between the DBM learned feature with it variants proposed in this paper.

Visual Feature	Reduction	Accuracy
DBM learned feature	None	59.9%
DBM learned feature	LDA	64.4%
DBM learned feature + DCT	LDA	69.1%

In order to demonstrate the superiority of the augmented visual feature over the DBM learned feature, we used a data visualisation technique called t-SNE [24] to produce 2D embeddings of the visual features. The points close in the high dimensional feature space are also close in the 2D space produced by t-SNE. Fig. 6 shows the 2D mapping of the augmented visual feature and the DBM learned feature. The points in Fig. 6 represent video frames, while the different colours correspond to different classes (i.e., different states of the HMM models). For clarity, we only randomly chose the fifth state of each digit HMM model for visualisation. Fig. 6 shows that, compared to the DBM learned feature, the augmented visual feature appears to be more visually discriminative than the DBM learned feature which exhibits more dispersion. This explains why the per-



(a)



(b)

Figure 6. 2D t-SNE visualization of the DBM learned feature (Fig. 6a) and our proposed feature (Fig. 6b).

formance of the augmented visual feature is better than the DBM learned feature.

We also compared our method with other VSR feature learning and extraction techniques, and listed the recognition accuracies in Table 4. In our experiments, we reported the speech accuracy obtained by the DCT visual features with LDA (for feature dimension reduction) which is a well-established framework and it has been used for decades for visual speech recognition. In addition to LDA, we also report the accuracy obtained by mutual information selectors [5] (i.e., MMI, mRMR and CMI), which are used as feature reduction techniques on both DCT and LBP-TOP features. For the hand-crafted features with the conventional feature reduction techniques, the best accuracy is achieved by the DCT feature with LDA (54.7%).

Regarding deep learning techniques, we compare our results with the stacked denoising auto-encoder introduced in

Table 4. Performance comparison between our proposed method with other feature learning and extraction techniques.

Feature Representation	Accuracy
DCT + LDA	54.7%
DCT + MMI [5]	52.3%
DCT + mRMR [5]	52.2%
DCT + CMI [5]	51.1%
LBP-TOP [29] + MMI [19]	52.5%
LBP-TOP [29] + mRMR [16]	53.1%
Deep Bottleneck Feature [23]	57.3%
Augmented Deep Bottleneck Feature [23]	67.8%
Our proposed method	69.1%

[23]. From Table 4, one can observe that the deep learning techniques generally perform better than the linear feature transformation and feature selectors. Meanwhile, our proposed method outperformed the augmented deep bottleneck visual feature which is based on the stacked denoising auto-encoder (69.1% vs 67.8%). Since the approximate inference of the DBM is performed in two directions (bottom-up and top-down), it is more capable of handling ambiguous inputs [18] compared to the stacked denoising auto-encoder.

5. Conclusion

In this paper, we propose a novel feature representation framework for lipreading. Unlike all the previous works which only use the visual information for both training and testing procedures, our work uses the audio information to augment the visual information and learn a better feature representation during the training phase. During the testing phase, the audio features can be inferred from the given visual information, and the inferred audio features can be further used along with the visual features to generate joint feature representations for lipreading. Experiments show that there is a significant accuracy improvement when using the multimodal DBM to learn the joint representation, since the missing audio information can be inferred to augment the feature representation. To the best of our knowledge, this is the first work that shows humans utterances can be reconstructed from humans lip movements. This novel framework provides a new solution to practical speech driven tasks where audio signals are corrupted by environmental noises.

References

[1] A. Bakry and A. Elgammal. Mkpls: Manifold kernel partial least squares for lipreading and speaker identification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 684–691. IEEE, 2013.

[2] L. Deng and R. Togneri. Deep dynamic models for learning hidden representations of speech features. In *Speech and Au-*

dio Processing for Coding, Enhancement and Recognition, pages 153–196. Springer, 2015.

[3] L. Deng and D. Yu. *Foundations and Trends in Signal Processing: DEEP LEARNING — Methods and Applications*. Microsoft Research, 2014.

[4] J. Gehring, Y. Miao, F. Metze, and A. Waibel. Extracting deep bottleneck features using stacked auto-encoders. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3377–3381. IEEE, 2013.

[5] M. Gurban and J.-P. Thiran. Information theoretic feature extraction for audio-visual speech recognition. *IEEE Transactions on Signal Processing*, 57(12):4765–4776, 2009.

[6] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 235–242. ACM, 2004.

[7] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[8] J. Huang and B. Kingsbury. Audio-visual deep learning for noise robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7596–7599. IEEE, 2013.

[9] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(2):198–213, 2002.

[10] K. Messer, J. Matas, J. Kittler, J. Lttin, and G. Maitre. XM2VTSDB: The extended m2vts database. In *In Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[11] V. Mnih. Cudamat: a cuda-based matrix class for python. *Department of Computer Science, University of Toronto, Tech. Rep. UTML TR*, 4, 2009.

[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.

[13] E. Ong and R. Bowden. Learning sequential patterns for lipreading. In *Proceedings of the 22nd British Machine Vision Conference*, 2011.

[14] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–2017. IEEE, 2002.

[15] Y. Pei, T.-K. Kim, and H. Zha. Unsupervised random forest manifold alignment for lipreading. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 129–136. IEEE, 2013.

[16] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[17] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004.

[18] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

[19] P. Scanlon and R. Reilly. Feature analysis for automatic speechreading. In *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*, pages 625–630. IEEE, 2001.

[20] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[21] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980, 2014.

[22] J. Su, A. Srivastava, F. D. de Souza, and S. 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.

[23] C. Sui, R. Togneri, and M. Bennamoun. Extracting deep bottleneck features for visual speech recognition. In *Proceedings of Acoustics, Speech, and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.

[24] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[25] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[26] M. Wagner, D. Tran, R. Togneri, P. Rose, D. Powers, M. Onslow, D. Loakes, T. Lewis, T. Kuratate, Y. Kinoshita, et al. The big Australian speech corpus (the big ASC). In *Proceedings of 13th Australasian International Conference on Speech Science and Technology*, pages 166–170, 2010.

[27] S. J. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK book version 3.4. 2006.

[28] D. Yu and M. L. Seltzer. Improved bottleneck features using pretrained deep neural networks. In *Proceedings of INTERSPEECH*, volume 237, page 240, 2011.

[29] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. *Multimedia, IEEE Transactions on*, 11(7):1254–1265, 2009.

[30] Z. Zhou, X. Hong, G. Zhao, and M. Pietikainen. A compact representation of visual speech data using latent variables. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1), Jan 2014.

[31] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen. A review of recent advances in visual speech decoding. *Image and Vision Computing*, 32(9):590–605, 2014.

[32] Z. Zhou, G. Zhao, and M. Pietikainen. Towards a practical lipreading system. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 137–144. IEEE, 2011.