Temporal Multimodal Learning in Audiovisual Speech Recognition

Di Hu*, Xuelong Li†, Xiaoqiang Lu†
*School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, P. R. China
†Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences,
Xi'an 710119, P. R. China
hdui831@mail.nwpu.edu.cn, xuelong_li@opt.ac.cn, luxiaoqiang@opt.ac.cn

Abstract

In view of the advantages of deep networks in producing useful representation, the generated features of different modality data (such as image, audio) can be jointly learned using Multimodal Restricted Boltzmann Machines (MRBM). Recently, audiovisual speech recognition based the MRBM has attracted much attention, and the MRBM shows its effectiveness in learning the joint representation across audiovisual modalities. However, the built networks have weakness in modeling the multimodal sequence which is the natural property of speech signal. In this paper, we will introduce a novel temporal multimodal deep learning architecture, named as Recurrent Temporal Multimodal RBM (RTMRBM), that models multimodal sequences by transforming the sequence of connected MRBMs into a probabilistic series model. Compared with existing multimodal networks, it's simple and efficient in learning temporal joint representation. We evaluate our model on audiovisual speech datasets, two public (AVLetters and AVLetters2) and one self-build. The experimental results demonstrate that our approach can obviously improve the accuracy of recognition compared with standard MRBM and the temporal model based on conditional RBM. In addition, RTMRBM still outperforms non-temporal multimodal deep networks in the presence of the weakness of long-term dependencies.

1. Introduction

Robust *Automatic Speech Recognition* (ASR) has been the key to the natural human-computer interfaces in most cases, but it's challenged by the noisy environments. One example of such an environment is street, where the traffic noise makes it very hard for recognizing the speech. Considering that vision is free of audio noise and can provide complemental information to audio in the noisy condition [10], even clean environment, researchers have paid attention to the *Audiovisual Speech Recognition* (AVSR) that makes use of the information from both audio and visual modalities. And the proposed several types of AVSR models have shown that they indeed have certain improvement over the ASR based on only audio [1, 10, 13].

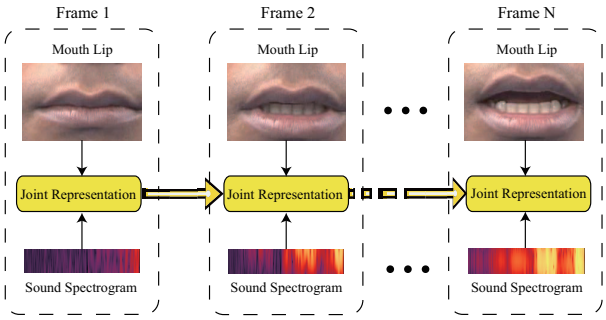

Figure 1: The sequence of audio and visual frames, where the joint representation across audiovisual modalities in current frame depends on the former. This directed graphical model is proposed to model temporal multimodal sequences in this paper.

In the past decades of years, several approaches have been proposed to fuse the speech information from the audio and visual modalities [5, 6, 14]. However, on account of the different statistical properties between the modalities [20], it's difficult to capture patterns across them. Recent works on deep learning [9, 16, 17] have verified the efficiency of deep networks in producing useful representations for various kinds of data, such as image, audio and text. It can be expected to explore the highly correlated representation across modalities after learning each channel data with single deep network. Based on this, multimodal deep networks have been proposed to jointly learn the generated features of different modalities and obtained state-of-

the-art performance [19, 20]. But many tasks are inherently sequential. For example, each utterance is an ordered sequence of phonemes or visemes (motions of mouth lips) in the AVSR, where the latter is influenced by the former. And the built multimodal networks almost fail to model the temporal multimodal data, which ignore the correlation among the components of the utterance.

In this paper, we propose a novel temporal multimodal network to model audiovisual sequence in an unsupervised fashion, which we refer to as the *Recurrent Temporal Multimodal Restricted Boltzmann Machine* (RTMRBM). Figure 1 shows a simple illustration of the proposed model. In each frame (time slice), the mouth lip and sound spectrogram are jointly learned using multimodal networks. The learned joint representations across modalities at different frames are directly connected from start to end, which makes the current frame learned based on previous one. In general, the proposed model has three advantages in the AVSR. First, it has the ability to extract semantic information from each modality data and learn the joint representation across audiovisual modalities. Second, it is a directed graphical model that can model temporal audiovisual sequences well as the joint representations among all frames are dependent. Third, the simple connection among the sequence of frames makes it easy to train as well as the standard MRBM. We evaluate our model on three audiovisual speech datasets, two public (AVLetters and AVLetters2) and one self-build (AVDigits). Our experiment results verify that the proposed model can learn better joint representation than non-temporal multimodal networks and temporal network based on *Conditional RBM* (CRBM). In addition, compared with typical multimodal network, RTMRBM can still performs well when faced with the weakness of long-term dependencies.

In the following sections, we first survey the related works about AVSR in Section 2. In Section 3, we review the representative multimodal model, then we develop the proposed RTMRBM, and introduce the inference as well as learning algorithm in it. Section 4 conducts different sets of experiments for evaluating the model on the three datasets, and corresponding results are reported and discussed. Section 5 concludes this paper.

2. Related Work

Classic AVSR Systems. AVSR has been studied in a few years, amounts of work about it can be roughly grouped into two categories: feature fusion and decision fusion [14]. The former aims to classify the concatenation of audio and visual features with a single classifier, but it has the weakness of separating out the noisy features. And the latter fuses the class-conditional probabilities of two classifiers with appropriate weights that depend on the contribution of each modality, such as multi-stream *Hidden Markov Models* (HMMs). However, The classic AVSR method based on multi-stream HMMs does not generalize very well because the weights that vary with time are hard to estimate. More importantly, both feature fusion and decision fusion have weakness in building a connection between audio and visual modalities at the level of semantics, where they are considered highly correlated [19].

AVSR based on Deep Learning. In recent years, deep learning methods have performed its effectiveness in generating useful feature representation. Most of the generated features from different kinds of data are considered as semantic correlated [20]. For the AVSR task, Ngiam *et al.* [13] proposed a kind of multimodal deep networks, *Multimodal Deep Autoencoder* (MDAE), which learns the layers of modality-specific network that consists stacks of RBMs firstly. Then, the joint representations across the generated features of audiovisual modalities are learned using *Multimodal RBM* (MRBM). Besides, the pre-trained MDAE is fine-tuned to minimize reconstruction errors of both modalities. Huang and Kingsbury [10] combined two *Deep Belief Networks* (DBNs) with the MRBM, and each DBN is used to model one type of modality. The organized *Multimodal DBN* (MDBN) has shown to outperform the accuracy of recognition by multi-stream HMMs. Similar frameworks have also been served to other tasks, such as multimodal retrieval [20]. MRBM has shown its ability in fusing the audio and visual modalities into a joint representation in the aforementioned networks. But the temporal information is not considered, which apparently deviates from the natural property of audiovisual speech signal. Recently, Amer *et al.* [1] attempted to model the audiovisual sequences for the first time. They made use of CRBM [25] to model each modality sequence in the task of AVSR, which made the modality-specific network sequence connected. Then the joint representation across modalities was generated. But, the multimodal network based on CRBM makes the MRBM complex, and it's difficult to learn the joint representation across multiple modalities because there're full connectivity among all the pairs of single modality layer and shared hidden layer [12].

3. The Proposed Model

In this work, our proposed model aims at fusing the temporal audio and visual representations into a joint representation sequence. In the following subsections, we first briefly review the MRBM model which is used to learn the joint representation across modalities. Then we introduce the RTMRBM model and explain the inference and learning procedure in it.

3.1. Multimodal Restricted Boltzmann Machine

The RBM is an undirected graphical model that defines a probability distribution of visible units using hidden unit-

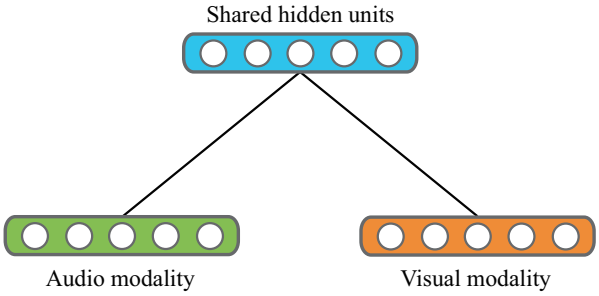

Figure 2: An illustration of the MRBM over audio and visual modality.

s [18]. Under the case of the multimodal input (we will take audiovisual inputs as an example), MRBM (Figure 2) defines the joint distribution over audio modality **a**, visual modality **v**, and shared hidden units **h** [19],

$$P(\mathbf{a}, \mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp\left(-E(\mathbf{a}, \mathbf{v}, \mathbf{h})\right), \tag{1}$$

where Z is the partition function and E is an energy function given by

$$\begin{aligned} E(\mathbf{a}, \mathbf{v}, \mathbf{h}) = &-\mathbf{a}^T \mathbf{W}^a \mathbf{h} - \mathbf{v}^T \mathbf{W}^v \mathbf{h} \\ &- \mathbf{a}^T \mathbf{b}^a - \mathbf{v}^T \mathbf{b}^v - \mathbf{h}^T \mathbf{b}^h, \end{aligned} \tag{2}$$

where **a** and **v** are the binary visible units of audio and visual input, and **h** is the binary shared hidden units. $\mathbf{W}^a$ is a matrix of pairwise weights between elements of **a** and **h**, and similar for $\mathbf{W}^v$. $\mathbf{b}^a$, $\mathbf{b}^v$, $\mathbf{b}^h$ are bias vectors for **a**, **v**, and **h**, respectively. To obtain the joint likelihood $P(\mathbf{a}, \mathbf{v})$, **h** is marginalized out from the distribution,

$$P(\mathbf{a}, \mathbf{v}) = \sum_{\mathbf{h}} \exp(-E(\mathbf{a}, \mathbf{v}, \mathbf{h}))/Z. \tag{3}$$

For the MRBM model, similar to the standard RBM, *Contrastive Divergence* (CD) [8, 23] or *Persistent CD* (PCD) [26] is used to approximate the gradient to maximize the joint likelihood, i.e., $P(\mathbf{a}, \mathbf{v})$. This is the typical maximum likelihood learning for MRBM. Finally, the learned shared hidden units **h** is treated as the joint representation across modalities.

3.2. Temporal Multimodal Learning

Although MRBM is good at learning the joint representation across modalities, it fails to capture the temporal information about the multimodal sequence, especially in AVSR. Specifically, the audio inputs of MDAE or MDBN are the concatenation of several frames of audio spectrogram, similarly for visual modality. These frames are only a part of each utterance representing phonemes or visemes, which ignores the continuity of all the frames that belong to the utterance. In addition, the joint representations obtained from multimodal networks are directly concatenated without considering the interaction and influence among them. To overcome the aforementioned problems, audiovisual representations should be viewed as sequence and modeled by temporal multimodal networks.

To model the audio and visual representation sequences simultaneously, it's intuitive to organize a sequence of MRBMs. The joint representations are considered to be connected among MRBMs, where the latter representations are dependent on the former. It's more credible than the generated representation of single modality [13], which can provide complement information for each modality. Researchers have also verified that merged information is more useful than the summation of single channel [21] in the field of cognitive science. In addition, the simple connections among MRBMs can identify the dependency correlation between joint and audiovisual layers, which means it can learn useful representation across modalities.

In fact, the organized network is a modification of *Temporal RBM* (TRBM) [24] which consists of a sequence of RBMs, where the hidden layer of current RBM depends on the previous RBMs. Although the TRBM has shown its effectiveness in modeling unimodal data, such as a sequence of bouncing ball or motion captures, it can not deal with the multimodal data.

3.3. Recurrent Temporal Multimodal RBM

The proposed RTMRBM models sequences of audio representation $\{\mathbf{a}_t\}_{t=1}^T$, video representation $\{\mathbf{v}_t\}_{t=1}^T$, where $\mathbf{a}_t \in \{0,1\}^{N_a}$, $\mathbf{v}_t \in \{0,1\}^{N_v}$, t is the time step and T is the sequence length. Figure 3 shows an illustration of the RTMRBM network. Specifically, the audio and visual representation $\{\mathbf{a}_t, \mathbf{v}_t\}$ form the aforementioned MRBM with the shared hidden units $\mathbf{h}_t \in \{0,1\}^{N_h}$ at time step t. Actually, it's hard to infer the shared hidden units $\mathbf{h}_t$ depended on former $\mathbf{h}_{t-1}$ exactly when the hidden layers of MRBMs are connected, because the required exact ratio of two MRBM partition functions is hard to evaluate [24]. Inspired by the recurrent TRBM [22], through making the connection between visual layer and the hidden layer directed and using mean-field update instead, exact inference becomes easy. Therefore, as described in Figure 3, we add the joint layers $\{\mathbf{J}_t\}_{t=1}^T$ on the top of MRBMs, which connects the sequence of MRBMs, where $\mathbf{J}_t \in R^{N_h}$.

The joint distribution over $\mathbf{a}_t$, $\mathbf{v}_t$, and $\mathbf{h}_t$ given the previous joint units $\mathbf{J}_{t-1}$ at time t is defined by the equation

$$P(\mathbf{a}_t, \mathbf{v}_t, \mathbf{h}_t | \mathbf{J}_{t-1}) = \frac{1}{Z_{J_{t-1}}} \exp\left(-E(\mathbf{a}_t, \mathbf{v}_t, \mathbf{h}_t | \mathbf{J}_{t-1})\right), \tag{4}$$

where $Z_{J_{t-1}}$ is the partition function that depends on J_{t-1},

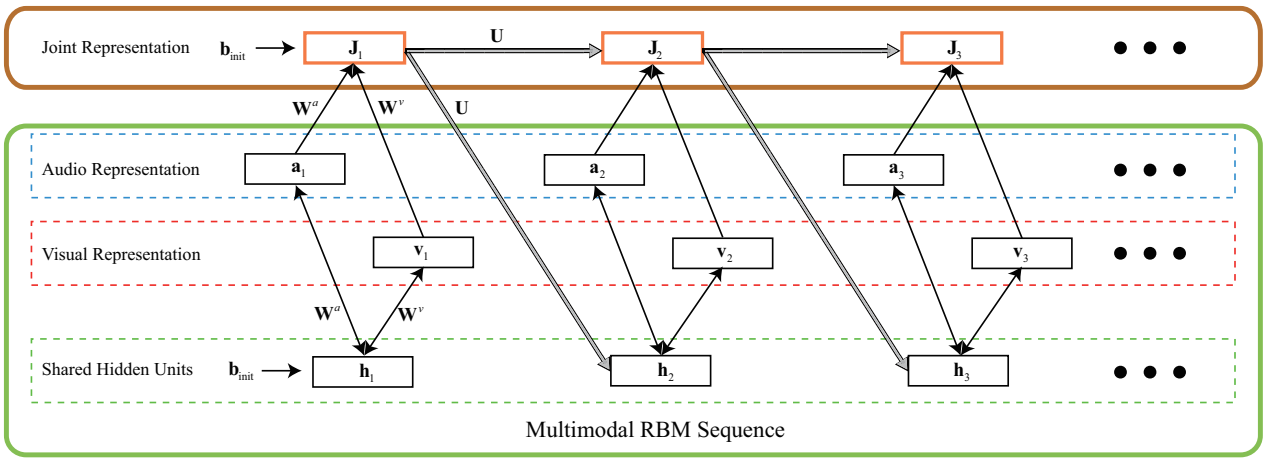

Figure 3: The structure of the RTMRBM network.

E is the energy function of the MRBM given by

$$\begin{aligned} E\left(\mathbf{a}_t, \mathbf{v}_t, \mathbf{h}_t | \mathbf{J}_{t-1}\right) &= -\mathbf{a}_t^T \mathbf{W}^a \mathbf{h}_t - \mathbf{v}_t^T \mathbf{W}^v \mathbf{h}_t \\ &\quad -\mathbf{a}_t^T \mathbf{b}^a - \mathbf{v}_t^T \mathbf{b}^v - \mathbf{h}_t^T \mathbf{b}^h \\ &\quad -\mathbf{h}_t^T \mathbf{U} \mathbf{J}_{t-1} \\ &= E\left(\mathbf{a}_t, \mathbf{v}_t, \mathbf{h}_t\right) - \mathbf{h}_t^T \mathbf{U} \mathbf{J}_{t-1}, \end{aligned} \tag{5}$$

where model parameters $\left\{\mathbf{W}^a, \mathbf{W}^v, \mathbf{b}^a, \mathbf{b}^v, \mathbf{b}^h\right\}$ are the matrixes of connection weights and biases for layers as in Eq.2. As for the matrix $\mathbf{U} \in R^{N_J \times N_J}$, it's about the pairwise weights of $\mathbf{J}_{t-1}$ and $\mathbf{J}_t$, which is the only difference compared with MRBM. When $t = 1$, $\mathbf{b}_{init}$ is treated as the input instead of term $(\mathbf{b}^h + \mathbf{U}\mathbf{J}_{t-1})$. Obviously, the energy function in Eq.5 consists of the energy of standard MRBM (Eq.2) and the term based on former joint units $\mathbf{J}_{t-1}$, which makes the RTMRBM model the multimodal sequence.

The mean-field value $\mathbf{J}_t$ is utilized to make the exact inference easier [12], which is essentially the expected value of $\mathbf{h}_t$ given the audio and visual representation $\{\mathbf{a}_t, \mathbf{v}_t\}$. Therefore, the sequence $\{\mathbf{J}_t\}_{t=1}^T$ is treated as the learned joint representation, and that is obtained as follows,

$$\mathbf{J}_t = \sigma\left(\mathbf{a}_t^T \mathbf{W}^a + \mathbf{v}_t^T \mathbf{W}^v + \mathbf{b}^h + \mathbf{U}\mathbf{J}_{t-1}\right), \tag{6}$$

where $\sigma(\cdot)$ is the element-wise logistic sigmoid function. When $t = 1$, $\mathbf{b}_{init}$ is used as before. Note that, Eq.6 makes the connected MRBM sequence into a kind of *Recurrent Neural Network* (RNN), where joint units $\mathbf{J}_t$ is time-dependent.

Given the former joint units $\mathbf{J}_{t-1}$, it's easy to obtain the joint probability over audio and visual representation $\{\mathbf{a}_t, \mathbf{v}_t\}$ by marginalizing out the hidden units $\mathbf{h}_t$,

$$P\left(\mathbf{a}_t, \mathbf{v}_t | \mathbf{J}_{t-1}\right) = \sum_{\mathbf{h}_t} \exp\left(-E\left(\mathbf{a}_t, \mathbf{v}_t, \mathbf{h}_t | \mathbf{J}_{t-1}\right)\right) / Z_{J_{t-1}}. \tag{7}$$

Based on the joint probability at time t, the joint probability over two sequences of audio and visual representation satisfies (see [12, 22] for more about the second equation),

$$\begin{aligned} \mathcal{F} &= \log P\left(\{\mathbf{a}_t\}_{t=1}^T, \{\mathbf{v}_t\}_{t=1}^T\right) \\ &= \sum_{t=1}^{T} \log P\left(\mathbf{a}_t, \mathbf{v}_t | \mathbf{J}_{t-1}\right). \end{aligned} \tag{8}$$

To model the audiovisual sequences, we'd like to maximize the joint likelihood $\mathcal{F}$.

3.4. Inference and Learning in RTMRBM

Inference in RTMRBM is achieved by giving previous joint units $\mathbf{J}_{t-1}$ and calculating the activation state of shared hidden layer $\mathbf{h}_t$ when $\mathbf{a}_t$ and $\mathbf{v}_t$ are fixed. Specifically, using the advantage of conditional independence in RBM, each unit in the shared hidden layer $\mathbf{h}_t$ is activated with the probability,

$$\begin{aligned} &P\left(\mathbf{h}_{t_k} = 1 | \mathbf{a}_t, \mathbf{v}_t, \mathbf{J}_{t-1}\right) \\ &= \sigma\left(\mathbf{a}_t^T \mathbf{W}_{\cdot k}^a + \mathbf{v}_t^T \mathbf{W}_{\cdot k}^v + \mathbf{b}_k^h + \mathbf{U}_{k\cdot} \mathbf{J}_{t-1}\right). \end{aligned} \tag{9}$$

Given joint representation sequence $\{\mathbf{J}_t\}_{t=1}^T$, the RTMRBM is decoupled into a sequence of MRBMs, and CD or PCD can be used for learning. Therefore, inferring the audiovisual representation $\{\mathbf{a}_t, \mathbf{v}_t\}$ is also required.

For the representation layers $\{\mathbf{a}_t, \mathbf{v}_t\}$ which both connect to the shared hidden layer $\mathbf{h}_t$, when $\mathbf{h}_t$ is observed, they cannot affect each other. Hence, the conditional distribution of units in $\mathbf{a}_t$ and $\mathbf{v}_t$ take the form as follows,

$$P\left(\mathbf{a}_{t_i} = 1 | \mathbf{v}_t, \mathbf{h}_t, \mathbf{J}_{t-1}\right) = \sigma\left(\mathbf{W}_{i\cdot}^a \mathbf{h}_t + \mathbf{b}_i^a\right), \tag{10}$$

$$P\left(\mathbf{v}_{t_j} = 1 | \mathbf{v}_t, \mathbf{h}_t, \mathbf{J}_{t-1}\right) = \sigma\left(\mathbf{W}_{j\cdot}^v \mathbf{h}_t + \mathbf{b}_j^v\right). \tag{11}$$

Compared with standard MRBM, the inference in RTMRBM is performed similarly except the joint units. The joint units come from previous time $t-1$ have impact on the activation of shared hidden units and audiovisual representation (indirectly) at time t. In other words, given $\mathbf{J}_{t-1}$, the learning of current MRBM depends on shared audiovisual representation and previous sequence, which makes the MRBM time-dependent.

Learning in the proposed model RTMRBM is performed by learning model parameters $\{\mathbf{W}^a, \mathbf{W}^v, \mathbf{b}^a, \mathbf{b}^v\}$ and $\{\mathbf{U}, \mathbf{b}^h, \mathbf{b}_{init}\}$. The former relates to standard MRBM, which we will focus on. The latter is related to joint units $\mathbf{J}_t$ and can be learned using the same learning rules as RTRBM [22], which is based on *Backpropagation Through Time* (BPTT) algorithm [15].

Similar with MRBM, parameters $\{\mathbf{W}^a, \mathbf{W}^v, \mathbf{b}^a, \mathbf{b}^v\}$ can be learned based on CD approximation but time-dependent. Specifically, $\mathbf{b}^a$ and $\mathbf{b}^v$ are learned as follows,

$$\mathbf{b}^a := \mathbf{b}^a + \alpha \sum_{t=1}^{T} \left(E_{P_{data}}\left[\mathbf{a}_t\right] - E_{P_{\text{recon}}}\left[\mathbf{a}_t\right] \right), \tag{12}$$

$$\mathbf{b}^v := \mathbf{b}^v + \alpha \sum_{t=1}^{T} \left(E_{P_{data}}\left[\mathbf{v}_t\right] - E_{P_{\text{recon}}}\left[\mathbf{v}_t\right] \right), \tag{13}$$

where α is a learning rate, $E_{P_{data}}$ is the data-dependent expectation, and $E_{P_{recon}}$ is the data-reconstruction's expectation but depends on the joint representation sequence $\{\mathbf{J}_t\}_{t=1}^{T-1}$.

$\mathbf{W}^a$ and $\mathbf{W}^v$ of $\mathcal{F}$ are updated using gradient ascent which consists of two terms, one is about inferring the joint representation $\mathbf{J}_t$ given audiovisual representation $\{\mathbf{a}_t, \mathbf{v}_t\}$, the other one relates to MRBM at time t given previous joint representation $\mathbf{J}_{t-1}$,

$$\begin{aligned} \Delta_{\mathbf{W}^a}\mathcal{F} = &\sum_{t=1}^{T-1} \mathbf{a}_t \left(\Delta_{\mathbf{J}_{t+1}}\mathcal{F} \odot \mathbf{J}_t \odot (\mathbf{1} - \mathbf{J}_t)\right)^T \\ &+ \sum_{t=1}^{T} \Delta_{\mathbf{W}^a} \log P\left(\mathbf{a}_t, \mathbf{v}_t | \mathbf{J}_{t-1}\right), \end{aligned} \tag{14}$$

$$\begin{aligned} \Delta_{\mathbf{W}^v}\mathcal{F} = &\sum_{t=1}^{T-1} \mathbf{v}_t \left(\Delta_{\mathbf{J}_{t+1}}\mathcal{F} \odot \mathbf{J}_t \odot (\mathbf{1} - \mathbf{J}_t)\right)^T \\ &+ \sum_{t=1}^{T} \Delta_{\mathbf{W}^v} \log P\left(\mathbf{a}_t, \mathbf{v}_t | \mathbf{J}_{t-1}\right), \end{aligned} \tag{15}$$

where $\odot$ denotes element-wise product. Note that the second term in Eq.14 is the summation over the negative gradient of MRBM at time t with regard to $\mathbf{W}^a$, which is computed as the standard MRBM using CD approximation $\left(E_{P_{data}}\left[\mathbf{a}_t\mathbf{h}_t^T\right] - E_{P_{\text{recon}}}\left[\mathbf{a}_t\mathbf{h}_t^T\right]\right)$, and similarly for the weight matrix $\mathbf{W}^v$ of visual modality (Eq.15). The term $\Delta_{\mathbf{J}_{t+1}}\mathcal{F}$ in both Eq.14 and Eq.15 takes the form

$$\begin{aligned} \Delta_{\mathbf{J}_t}\mathcal{F} = &\mathbf{U}^T \left(\mathbf{J}_{t+1} \odot (\mathbf{1} - \mathbf{J}_{t+1}) \odot \Delta_{\mathbf{J}_{t+1}}\mathcal{F}\right) \\ &+ \mathbf{U}^T \Delta_{\mathbf{b}^h} \log P\left(\mathbf{a}_t, \mathbf{v}_t | \mathbf{J}_{t-1}\right). \end{aligned} \tag{16}$$

$\Delta_{\mathbf{J}_t}\mathcal{F}$ is computed recursively, and $\mathbf{J}_{T+1} = 0$. The term $\Delta_{\mathbf{b}^h} \log P\left(\mathbf{a}_t, \mathbf{v}_t | \mathbf{J}_{t-1}\right)$ is also computed with CD approximation. We summarize the learning procedure in Algorithm 1.

Algorithm 1 Learning in RTMRBM

Input: Audio representation $\{\mathbf{a}_t\}_{t=1}^T$, Visual representation $\{\mathbf{v}_t\}_{t=1}^T$, CD steps K, number of iteration N.
Output: Model parameters $\{\mathbf{W}^a, \mathbf{W}^v, \mathbf{b}^a, \mathbf{b}^v, \mathbf{b}^h, \mathbf{b}_{init}\}$.

1: Initialize model parameters.
2: Compute the joint representation sequence $\{\mathbf{J}_t\}_{t=1}^T$ using Eq.6.
3: Run a Gibbs chain for K steps for each MRBM given $\mathbf{J}_{t-1}$ at time step t, the sampling probability for $\mathbf{b}_t$, $\mathbf{a}_t$, and $\mathbf{v}_t$ follows Eq.9, Eq.10, and Eq.11, respectively.
4: Update the bias of audio and visual layers $\{\mathbf{b}^a, \mathbf{b}^v\}$ follows the Eq.12 and Eq.13. The pairwise matrixes $\mathbf{W}^a$ and $\mathbf{W}^v$ are updated according to the rules $\mathbf{W}^a := \mathbf{W}^a + \alpha\Delta_{\mathbf{W}^a}\mathcal{F}$ and $\mathbf{W}^v := \mathbf{W}^v + \alpha\Delta_{\mathbf{W}^v}\mathcal{F}$, respectively.
5: Update $\{\mathbf{U}, \mathbf{b}^h, \mathbf{b}_{init}\}$ using the learning rules of RTRBM [22].
6: Repeat above 2-5 steps until convergence or N steps

4. Experiments

In this section, we show the results of RTMRBM compared with other models on three datasets, including MDAE, MDBN, and CRBM. Different *Signal Noisy Ratio* (SNR) are also added to audio signal for evaluating the performance. In addition, we analyze the impact of the joint-joint weight matrix $\mathbf{U}$ on sub-sequence when modeling the whole multimodal sequence.

4.1. Datasets

Experiments are conducted on three datasets, two public datasets: AVLetters [11], AVLetters2 [6], and one self-build dataset: AVDigits. The AVDigits dataset is built to examine the performance of RTMRBM, which contains different semantic information compared with the other two.

AVLetters contains 10 speakers speaking the letters A to Z at three times each. This dataset provides pre-extracted lip regions of 60×80 pixels and audio features (raw audio is not provided) *Mel-Frequency Cepstrum Coefficient* (MFCC). Similar with [1], the training set of this dataset contains the first two times of each letter spoken by each speaker,

and the rest is for the test set. Hence, the training set and testing set both contain the same set of speakers, which is speaker dependent.

AVLetters2 is a high-definition of AVLetters. It's about reading letters from A to Z, spoke by five people, seven times for each letter. Similar with [6], letters spoken by four people are for training and the rest one is for testing. This is different from previous train/test split, which is speaker independent.

AVDigits is a self-build dataset, which is about speaking digits. We ask 6 people to face the camera and speak digits 0 to 9 at nine times each. All videos are recorded in full-frontal pose, and all subjects are required to keep their heads fixed as far as possible. Speakers are also asked to close their mouth at the begin and end when speaking prepared digits. The recording video devices is SONY cx290, the visual modality of each utterance is digitized in 1920×1080 at 25fps, and audio is recorded at 48kHz, 16-bit resolution. Letters spoken by four people are exploited to train and the rest two are exploited to test, which is also speaker independent.

4.2. Data Preprocessing

The audio and visual data are separated and preprocessed, respectively. For audio signal, spectrogram (MFCC instead in AVLetters) is extracted with 20ms hamming window and 10ms overlap. The frequency points of Discrete Fourier Transforms are 500, which results in 251 dimension vector of the signal window. The obtained spectral coefficient vector is reduced to 50 dimensions using PCA whitening.

For visual signal, cascade object detector [27] is used to extract the Region-of-Interest that encompasses the mouth. The extracted region is rescaled to 60×80 pixels and reduced to 100 principal components using PCA whitening as well. We use 4 contiguous audio frames and 1 video frames as the inputs for each time step simultaneously, which are almost the same duration.

4.3. Implementation Details

As *Deep Auto-encoder* (DAE) can be used to obtain the efficient binary codes for both audio and visual information [7, 9], the audiovisual representations are pre-extracted using modality-specific DAEs. In addition, lots of experiments have verified that pretraining method can indeed affect the performance of RNN [4]. We find that the initialized $\mathbf{W}^a$, $\mathbf{W}^v$, $\mathbf{b}^a$, $\mathbf{b}^v$, and $\mathbf{b}^h$ from MDAE can capture more shared information across audiovisual modalities. The pairwise weight $\mathbf{U}$ and initial value $\mathbf{b}_{init}$ can be initialized to small random value.

For the task of AVSR, the joint representation generated from the RTMRBM is treated as the audiovisual fusion, which can be learned in an unsupervised manner. Since each speaking example has varying duration, we divided the fusion result into 1 and 3 equal slices, similar to [13]. Each slice consists of several audio and visual frames, and mean-pooling is performed over them. Then, the obtained features of each slice are concatenated and classified using a linear SVM.

4.4. Results

To evaluate the joint representation learned by our proposed RTMRBM model, we conduct sets of experiments on both unimodal and multimodal data, i.e. audio modality, visual modality, and both of them. For the unimodal fashion, we present only one modality and set the other one to be zero during the learning procedure.

4.4.1 Speaker Dependent

In the speaker dependent experiments, we compare RTMRBM with the performance of several methods on AVLetters, which includes prior methods based on hand-craft features and multimodal deep networks. Table.1 shows the comparison results of mean accuracy over all the letters. For the three kinds of modalities, RTMRBM outperforms all the others. In contrast, HMM is also trained to model temporal sequence with 3DCNN, but RTMRBM deals with the visual data simpler and learns better joint representation. In addition, RTMRBM has an improvement compared with non-temporal multimodal deep networks MDAE that is based on MRBM, which means our proposed model learns better feature representation through capturing the temporal information. Note that, for the audiovisual modalities, CRBM makes the modality-specific network connected instead of the joint representation which can capture bet-

Modality	Model	mean Accuracy
A	MDAE [13]	58.40
	CRBM [1]	61.2
	RTMRBM	**64.41**
V	Multiscale Spatial Analysis [11]	44.6
	Local Binary Pattern [29]	58.85
	3DCNN-HMM [28]	59.6
	MDAE [13]	62.10
	CRBM [1]	62.60
	RTMRBM	**64.63**
AV	MDAE [13]	62.90
	CRBM [1]	64.8
	RTMRBM	**66.04**

Table 1: The mean accuracy of speech classification on AVLetters, RTMRBM and other models are evaluated with single audio/visual modality and both of them.

Dataset	Model	Modality	SNR					Clean
			-4dB	4dB	6dB	10dB	12dB	
AVLetters2	AAM [2]	V	15.2	15.2	15.2	15.2	15.2	15.2
	RTMRBM	V	31.21	31.21	31.21	31.21	31.21	31.21
		A	42.25	57.01	58.04	62.31	67.71	**75.85**
		AV	**56.66**	**64.20**	**63.02**	**66.80**	**69.66**	74.77
	MDAE [13]	AV	49.61	60.22	62.21	64.13	66.31	67.89
	MDBN [10]	AV	43.57	44.12	46.87	47.52	49.07	54.10
AVDigits	RTMRBM	V	40.66	40.66	40.66	40.66	40.66	40.66
		A	43.50	54.09	57.67	62.02	64.42	71.02
		AV	**55.36**	**61.52**	**62.02**	**64.11**	**67.64**	**71.77**
	MDAE [13]	AV	51.57	58.96	59.52	61.63	64.52	66.74
	MDBN [10]	AV	49.44	50.00	52.78	53.33	54.44	55.00

Table 2: Speech classification performance on AVLetters2 and AVDigits. The results show that, the RTMRBM performs better than MDAE and MDBN under the conditions of different degrees of SNR to the audio signal, also almost better than single modality.

ter features across modalities, therefore it's hard for CRBM to learn better joint representation based on previous one. These classification results show the efficient of RTMRBM in generating feature representation on both single modality and multi-modalities.

4.4.2 Speaker Independent

In the speaker independent experiments, we compare RTMRBM mainly with MDAE and MDBN models on AVLetters2 and AVDigits. To evaluate our proposed model at different levels of audio noise, we add the white Gaussian noise from -4dB to 12dB SNR to the original clean signal. Table.2 shows the comparison among modalities and models. There're three points we should pay attention to. First, on the AVLetters2 dataset, we make a contrast with *Active Appearence Model* (AAM) [2] on the visual modality, which learns a mean face template, but it's sensitive to the specific training speakers. The results show that modeling sequence with RTMRBM lower the degree of sensitivity to some extent. Second, on both datasets, we indeed improve the mean accuracy at different levels of SNR by learning both audio and visual modalities instead of one of them. Especially, when audio SNR becomes lower, the accuracy of audiovisual modality has a significant improvement compared with single modality, which ensures the learned joint representation sequence has more discriminative and robust features. We also note that the audiovisual modality performs worse than single audio information in the situation of clean audio signal. This is because the visual modality lower the performance, which is a common situation [13]. Third, compared with the other multimodal deep networks (MDAE and MDBN), RTMRBM performs better on both datasets. This shows that modeling the multimodal sequences indeed capture temporal information and therefore make the model learn better feature representation across modalities.

4.5. Additional Contrast Experiment

As the RTMRBM is essentially a RNN, it's challenged by the long-term dependencies [3]. In this experiment, through examining the effectiveness of the proposed model on sub-sequence when modeling the whole audiovisual modalities sequence, we explore the degree of influence. Specifically, the RTMRBM is trained with the whole sequence of speaking examples, but the generated joint representation is equally divided into three parts, former, middle and latter. Feature representation in each part are trained

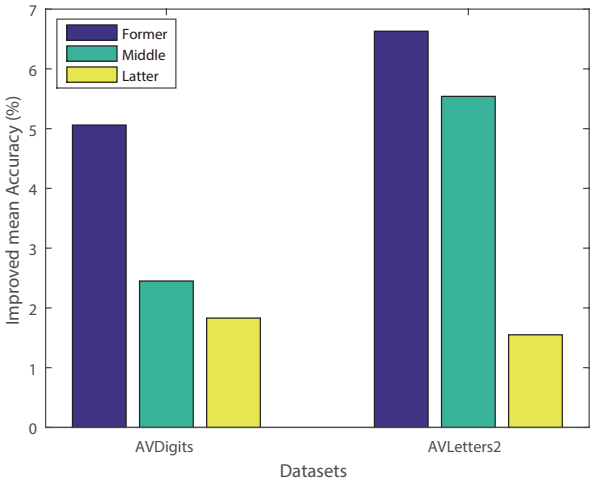

Figure 4: The contrast among the improved mean accuracy of three parts (former, middle, and latter) of utterances, compared with MDAE on AVDigits and AVLetters2.

and classified with a linear SVM. And then we make a comparison with MDAE which is treated as the same fashion. Figure 4 shows the improved mean accuracy of each part compared with MDAE on AVDigits and AVLetters2. The results shows that RTMRBM learns better feature representation on all the three sequences compared with MDAE, but they have different degree of improvement. On both datasets, the former part of sequence is enhanced more than the latter, which ensures that RTMRBM indeed has difficulty in tackling the long-term dependencies. However, through modeling the multimodal sequence, the temporal information plays an important role in learning the joint representation. RTMRBM can still learn better joint feature representation on the latter part than non-temporal multimodal network.

5. Conclusion

We have proposed a new architecture RTMRBM for modeling temporal multimodal sequences, which makes the sets of MRBMs model the temporal multimodal data well because the previous robust joint representation is provided for the learning of current MRBM. Meanwhile, the simple connection makes the model easy to train. Our experimental results show that the proposed model can learn temporal joint representation across multiple modalities in the task of AVSR, even when the different levels of audio noise exist. In addition, although the model is truly affected by joint-joint weight matrix in the long-term dependency, it also performs better than non-temporal multmodal networks, which verifies the importance of temporal information and its effectiveness. In the future, we plan to apply the RTMRBM in other temporal multimodal tasks and attempt to reduce the difficulties in learning long-term multimodal sequences.

6. Acknowledgement

This work is supported by the Fundamental Research Funds for the Central Universities (Grant no. 3102015BJ(II)JJZ01), the National Basic Research Program of China (973 Program) (Grant No. 2012CB719905), State Key Program of National Natural Science of China (Grant No. 61232010), the National Natural Science Foundation of China (Grant No. 61472413), and by the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences (Grant No. LSIT201408).

References

[1] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney. Multimodal fusion using dynamic hybrid models. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 556–563. IEEE, 2014.

[2] H. L. Bear, S. J. Cox, and R. W. Harvey. Speaker-independent machine lip-reading with speaker-dependent viseme classifiers. 2015.

[3] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5(2):157–166, 1994.

[4] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv preprint arXiv:1206.6392*, 2012.

[5] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[6] S. J. Cox, R. Harvey, Y. Lan, J. L. Newman, and B.-J. Theobald. The challenge of multispeaker lip-reading. In *AVSP*, pages 179–184. Citeseer, 2008.

[7] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. E. Hinton. Binary coding of speech spectrograms using a deep auto-encoder. In *Interspeech*, pages 1692–1695. Citeseer, 2010.

[8] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

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

[10] 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.

[11] 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.

[12] R. Mittelman, B. Kuipers, S. Savarese, and H. Lee. Structured recurrent temporal restricted boltzmann machines. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1647–1655, 2014.

[13] 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*, pages 689–696, 2011.

[14] 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.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5:3, 1988.

[16] R. Salakhutdinov and G. Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.

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

[18] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. 1986.

[19] K. Sohn, W. Shang, and H. Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014.

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

[21] B. E. Stein and M. A. Meredith. *The merging of the senses.* The MIT Press, 1993.

[22] I. Sutskever, G. E. Hinton, and G. W. Taylor. The recurrent temporal restricted boltzmann machine. In *Advances in Neural Information Processing Systems*, pages 1601–1608, 2009.

[23] I. Sutskever and T. Tieleman. On the convergence properties of contrastive divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 789–795, 2010.

[24] G. W. Taylor, G. E. Hinton, and S. T. Roweis. Modeling human motion using binary latent variables. In *Advances in neural information processing systems*, pages 1345–1352, 2006.

[25] G. W. Taylor, G. E. Hinton, and S. T. Roweis. T-wo distributed-state models for generating high-dimensional time series. *The Journal of Machine Learning Research*, 12:1025–1068, 2011.

[26] T. Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM, 2008.

[27] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[28] D. Wu. *Human Action Recognition Using Deep Probabilistic Graphical Models*. PhD thesis, University of Sheffield, 2014.

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