# Multimodal Deep Learning

**Jiquan Ngiam**[1], **Aditya Khosla**[1], **Mingyu Kim**[1], **Juhan Nam**[2], **Honglak Lee**[3], **Andrew Y. Ng**[1]
[1] Computer Science Department, Stanford University
{jngiam,aditya86,minkyu89,ang}@cs.stanford.edu
[2] Department of Music, Stanford University
juhan@ccrma.stanford.edu
[3] Computer Science & Engineering Division, University of Michigan, Ann Arbor
honglak@eecs.umich.edu

## Abstract

Deep networks have been successfully applied to unsupervised feature learning for single modalities (e.g., text, images or audio). In this work, we propose a novel application of deep networks to learn features over multiple modalities. We present a series of tasks for multimodal learning and show how to train a deep network that learns features to address these tasks. In particular, we demonstrate cross modality feature learning, where better features for one modality (e.g., video) can be learned if multiple modalities (e.g., audio and video) are present at feature learning time. Furthermore, we show how to learn a shared representation between modalities and evaluate it on a unique task, where the classifier is trained with audio-only data but tested with video-only data and vice-versa. We validate our methods on the CUAVE and AVLetters datasets with an audio-visual speech classification task, demonstrating superior visual speech classification on AVLetters and effective multimodal fusion.

## 1 Introduction

In speech recognition, people are known to integrate audio-visual information in order to understand speech. This was first exemplified in the McGurk effect [1] where a visual */ga/* with a voiced */ba/* is perceived as */da/* by most subjects. In particular, the visual modality provides information on the place of articulation [2] and muscle movements which can often help to disambiguate between speech with similar acoustics (e.g., the unvoiced consonants */p/* and */k/*). In this paper, we examine multimodal learning and how to employ deep architectures to learn multimodal representations.

Multimodal learning involves relating information from multiple sources. For example, images and 3-d depth scans are correlated at first-order as depth discontinuities often manifest as strong edges in images. Conversely, audio and visual data for speech recognition have non-linear correlations at a "mid-level", as phonemes or visemes; it is difficult to relate raw pixels to audio waveforms or spectrograms.

In this paper, we are interested in modeling "mid-level" relationships, thus we choose to use audio-visual speech classification to validate our methods. In particular, we focus on learning representations for speech audio which are coupled with videos of the lips.

We will consider the learning settings shown in Figure 1. The overall task can be divided into three phases – feature learning, supervised training, and testing. We keep the supervised training and testing phases fixed and examine different feature learning models with multimodal data. In detail, we consider three learning settings – multimodal fusion, cross modality learning, and shared representation learning.

| | Feature Learning | Supervised Training | Testing |
|---|---|---|---|
| Classic Deep Learning | Audio | Audio | Audio |
| | Video | Video | Video |
| Multimodal Fusion | Audio + Video | Audio + Video | Audio + Video |
| Cross Modality Learning | Audio + Video | Audio | Audio |
| | Audio + Video | Video | Video |
| Shared Representation Learning | Audio + Video | Audio | Video |
| | Audio + Video | Video | Audio |

Figure 1: Multimodal Learning Settings.

For the multimodal fusion setting, data from all modalities is available at all phases; this represents the typical setting considered in most prior work in audio-visual speech recognition [3]. In cross modality learning, one has access to data from multiple modalities only during feature learning. During the supervised training and testing phase, only data from a single modality is provided. In this setting, the aim is to learn better single modality representations given unlabeled data from multiple modalities. Last, we consider a shared representation learning setting, which is unique in that different modalities are presented for supervised training and testing. This setting allows us to evaluate if the feature representations can capture correlations across different modalities. Specifically, studying this setting allows us to assess whether the learned representations are modality-invariant.

In the following sections, we first describe the building blocks of our model. We then present different multimodal learning models leading to a deep network that is able to perform the various multimodal learning tasks. Finally, we report experimental results and conclude.

## 2 Background

The multimodal learning settings we consider can be viewed as a special case of self-taught learning [4]. The self-taught learning paradigm uses unlabeled data (not necessarily from the same distribution as the labeled data) to learn representations that improve performance on some task. While self-taught learning was first motivated with sparse coding, recent work on deep learning [5, 6, 7] have examined how deep sigmoidal networks can be trained to produce useful representations for handwritten digits and text. The key idea is to use greedy layer-wise training with Restricted Boltzmann Machines (RBMs) followed by fine-tuning. We use an extension of RBMs with sparsity [8], which have been shown to be able to learn meaningful features for digits and natural images. In the next section, we review the sparse RBM, which we use as a layer-wise building block for our models.

### 2.1 Sparse restricted Boltzmann machines

We first describe the restricted Boltzmann machine (RBM) [5, 6] followed by the sparsity regularization method [8]. The RBM is an undirected graphical model with hidden variables ($\mathbf{h}$) and visible variables ($\mathbf{v}$). There are symmetric connections between the hidden and visible variables ($w_{i,j}$), but no connections between hidden variables or visible variables. This particular configuration makes it easy to compute the conditional probability distributions, when $\mathbf{v}$ or $\mathbf{h}$ is fixed (Equation 2).

$$-\log P(\mathbf{v},\mathbf{h}) \propto E(\mathbf{v},\mathbf{h}) = \frac{1}{2\sigma^2}\sum_i v_i^2 - \frac{1}{\sigma^2}\left(\sum_i c_i v_i + \sum_j b_j h_j + \sum_{i,j} v_i h_j w_{i,j}\right) \quad (1)$$

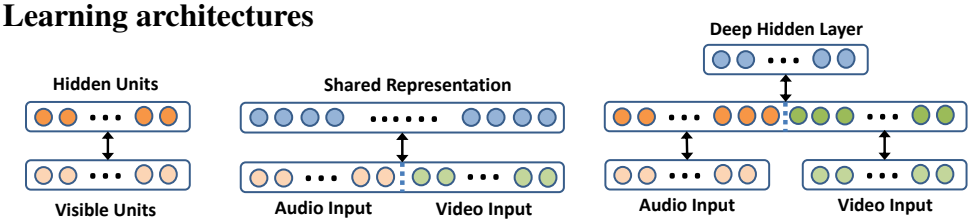
$$p(h_j|\mathbf{v}) = sigmoid(\tfrac{1}{\sigma^2}(b_j + \mathbf{w}_j^T\mathbf{v})) \quad (2)$$

Equation 1 gives the negative log-probability of a RBM while Equation 2 gives the posteriors of the hidden variables given the visible variables. This formulation models the visible variables as real-valued units and the hidden variables as binary units.[1] As it is intractable to compute the gradient of the log-likelihood term, we learn the parameters of the model ($w_{i,j}, b_j, c_i$)

[1]We use Gaussian visible units for the RBM that is connected to the input data. When training the deeper layers, we use binary visible units.

using contrastive divergence [9]. To regularize the model for sparsity, we encourage each hidden unit to have a pre-determined expected activation using a regularization penalty of the form $\lambda \sum_j (\rho - \frac{1}{m}(\sum_{k=1}^{m} \mathbf{E}[h_j|\mathbf{v}^k]))^2$, where $\{\mathbf{v}^1, ..., \mathbf{v}^m\}$ is the training set and $\rho$ determines the sparseness of the hidden units.

## 3 Learning architectures




(a) Standard RBM (b) Shallow RBM (c) Deep RBM

Figure 2: RBM Pretraining Models. We train (a) for audio and video separately as a baseline. The shallow model (b) is limited and we find that this model is unable to capture correlations across the modalities. The deep model (c) is trained in a greedy layer-wise fashion by first training two separate (a) models. We later "unroll" the deep model (c) to train the deep autoencoder models presented in Figure 3.

In this section, we describe our models for the task of audio-visual bimodal feature learning, where the audio and visual input to the model are windows of audio (spectrogram) and video frames. To motivate our deep autoencoder [5] model, we first describe several simple models and their drawbacks.

One of the most straightforward approaches to feature learning is to train a RBM model *separately* for audio and video (Figure 2a). After learning the RBM, the posteriors of the hidden variables given the visible variables (Equation 2) can then be used as a new representation for the data. We use this model as a baseline to compare the results of our multimodal learning models, as well as for pre-training the deep networks.

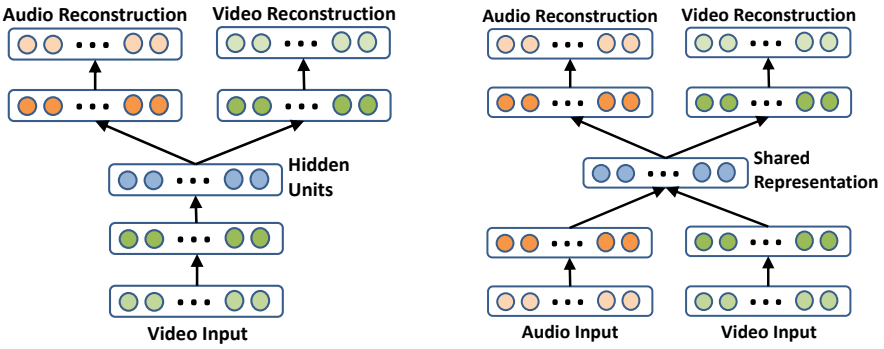
To train a multimodal model, an direct approach is to train a RBM over the concatenated audio and video data (Figure 2b). While this approach jointly models the distribution of the audio and video data, it is limited as a shallow model. In particular, since the correlations between the audio and video data are highly non-linear, it is hard for a RBM to learn these correlations and form multimodal representations.

Therefore, we consider greedily training a RBM over the pre-trained layers for each modality, as motivated by deep learning methods (Figure 2c). In particular, the posteriors (Equation 2) of the first layer hidden variables are used as the training data for the new layer. By essentially representing the data through learned first layer representations, it can be easier for the model to learn the higher-order correlations across the modalities. Intuitively, the first layer representations correspond to phonemes and visemes (lip pose and motions) and the second layer models the relationships between them.

However, there are still two issues with the above multimodal models. First, there is no explicit objective for the models to discover correlations across the modalities. It is possible for the model to find representations such that some hidden units are tuned only for audio while others are tuned only for video. Second, the models are clumsy to use in a cross modality learning setting where only one modality is present during supervised training and testing time. To use the RBM models presented above with only a single modality present, one would need to integrate out the other unobserved visible variables to perform inference.

Thus, we propose an autoencoder-based model that resolves both issues for the cross modality learning setting. The deep autoencoder (Figure 3a) is trained to reconstruct both modalities when given only video data. We initialize the deep autoencoder with the deep RBM weights (Figure 2c) based on Equation 2, discarding any weights that are no longer present due to the network's configuration. The middle layer is used as the new feature representation. This model can be viewed as an instance of multitask learning [10].

We use the deep autoencoder (Figure 3a) models in settings where only a single modality is present at supervised training and testing. On the other hand, when multiple modalities are available at




(a) Video-Only Deep Autoencoder

(b) Bimodal Deep Autoencoder

Figure 3: Deep Autoencoder Models. A "video-only" model is shown in (a) where the model learns to reconstruct both modalities given only video as the input. A similar model can be drawn for the "audio-only" setting. We train the (b) bimodal deep autoencoder in a denoising fashion, using an augmented dataset with examples that require the network to reconstruct both modalities given only one. Both models are pre-trained using sparse RBMs (Figure 2c). Since we use a sigmoid transfer function in the deep network, we can initialize the network using the conditional probability distributions $p(\mathbf{h}|\mathbf{v})$ and $p(\mathbf{v}|\mathbf{h})$ of the learned RBM.

task time, it is less clear how to use the model as one would need to train a deep autoencoder for each modality. One straightforward solution is to train the networks such that the decoding weights are tied. However, such an approach does not scale well – if we were to allow any combination of modalities to be present or absent at test time, we will need to train an exponential number of models. Instead, we propose a training method inspired by denoising autoencoders [11].

We propose training the deep autoencoder network (Figure 3b) using an augmented dataset with additional examples that have only a single-modality as input. In practice, we add examples that zero out one of the input modalities (e.g., video) and only have the other input modality (e.g., audio) available, yet still requiring the network to reconstruct both modalities (audio and video). Thus, one-third of the training data has only video for input, while another one-third of the data has only audio for input, and the last one-third of the data has both audio and video for input.

Due to initialization using sparse RBMs, we find that the hidden units have low expected activation even after the deep autoencoder training. Therefore, when one of the modalities is set to zero, the first layer representations are close to zero. In this case, we are essentially training a modality-specific deep autoencoder network (Figure 3a). Effectively, the method learns a model which is robust to missing modalities.

## 4 Experiments

We evaluate our methods on audio-visual speech classification of isolated letters and digits. The sparseness parameter $\rho$ was chosen using cross-validation, while all other parameters (including hidden layer size and weight regularization) were kept fixed.[2]

### 4.1 Data Preprocessing

We represent the audio signal using its spectrogram[3] with temporal derivatives, resulting in a 483 dimension vector which was reduced to 100 dimensions with PCA whitening. A window of 10 contiguous audio frames was used as the input to our models.

[2]We cross-validated $\rho$ over $\{0.01, 0.03, 0.05, 0.07\}$. The first layer features was 4x overcomplete for video (1536 units) and 1.5x overcomplete for audio (1500 units). The second layer had 4554 units.

[3]Each spectrogram frame (161 frequency bins) had a 20ms window with 10ms overlaps.

For the video, we preprocessed the frames so as to extract only the region-of-interest (ROI) encompassing the mouth.[4] Each mouth ROI was rescaled to 60x80 pixels and further reduced to 32 dimensions,[5] using PCA whitening. Temporal derivatives were computed over the reduced vector. We use windows of 4 contiguous video frames for input since this had approximately the same duration as 10 audio frames.

For both modalities, we also performed feature mean normalization over time [3], akin to removing the DC component from each example. We also note that adding temporal derivatives to the representations has been widely used in the literature as it helps to model dynamic speech information [3, 14]. The temporal derivatives were computed using a normalized linear slope so that the dynamic range of the derivative features are comparable to the original signal.

### 4.2 Datasets and Task

Since only unlabeled data was required for unsupervised feature learning, we combined diverse datasets to learn features. We used all the datasets for feature learning. AVLetters and CUAVE were further used for supervised classification. We ensured that no test data was used for unsupervised feature learning.

**CUAVE** [15]. 36 individuals saying the digits *0* to *9*. We used the *normal* portion of the dataset where each speaker was frontal facing and spoke each digit 5 times. We evaluated digit classification on the CUAVE dataset in a speaker independent setting. As there has not been a fixed protocol for evaluation on this dataset, we chose to use odd-numbered speakers for the test set and even-numbered ones for the training set.

**AVLetters** [16]. 10 speakers saying the letters *A* to *Z*, three times each. The dataset provided pre-extracted lip regions at 60x80 pixels. As we were not able to obtain the raw audio information for this dataset, we used it for evaluation on a visual-only lipreading task. We report results on the *third-test* settings used by [14, 16] for comparisons.

**AVLetters2** [17]. 5 speakers saying the letters *A* to *Z*, seven times each. This is a new high definition version of the AVLetters dataset. We used this dataset for unsupervised training only.

**Stanford Dataset**. 23 volunteers spoke the digits *0* to *9*, letters *A* to *Z* and selected sentences from the TIMIT dataset. We collected this data in a similar fashion to the CUAVE dataset and used for unsupervised training only.

**TIMIT**. We used the TIMIT [18] dataset for unsupervised audio feature pre-training.

We note that in all datasets there is variability in the lips in terms of appearance, orientation and size.

Our features were evaluated on speech classification of isolated letters and digits. We extracted features from overlapping windows. Since examples had varying durations, we divided each example into $S$ equal slices and performed average-pooling over each slice. The features from all slices were subsequently concatenated together. We combined features using $S = 1$ and $S = 3$ to form our final feature representation for classification using a linear SVM.

### 4.3 Cross Modality Learning

We first evaluate the learned features in a setting where unlabeled data for both modalities are available during feature learning, while during supervised training and testing phases only a single modality is presented. In these experiments, we evaluate cross modality learning where one learns better representations for one modality (e.g., video) when given multiple modalities (e.g., audio and video) during feature learning. For the bimodal deep autoencoder, we set the value of the other modality to zero when computing the shared representation which is consistent with the feature learning phase. All deep autoencoder models are trained with all available unlabeled audio and video data.

On the AVLetters dataset (Table 1a), there is an improvement over hand-engineered features from prior work. The deep autoencoder models performed the best on the dataset, obtaining a classification score of 65.8%, outperforming the best previous published results.

[4] We used an off-the-shelf object detector [12] with median filtering over time to extract the mouth regions.

[5] Similar to [13] we found that 32 dimensions were sufficient and performed well.

| Feature Representation | Accuracy |
|---|---|
| Baseline Preprocessed Video | 46.2% |
| RBM Video | 53.1% |
| Bimodal Deep Autoencoder | 59.2% |
| **Video-Only Deep Autoencoder** | **65.8%** |
| Multiscale Spatial Analysis [16] | 44.6% |
| Local Binary Pattern [14] | 58.9% |

(a) AVLetters

| Feature Representation | Accuracy |
|---|---|
| Baseline Video | 58.5% |
| RBM Video | 65.5% |
| Bimodal Deep Autoencoder | 66.7% |
| **Video-Only Deep Autoencoder** | **69.7%** |
| Discrete Cosine Transform [19] | 64% †§ |
| Active Appearence Model [20] | 75.7% † |
| Active Appearence Model [21] | 68.7% † |
| Fused Holistic+Patch [22] | 77.1% † |
| Visemic AAM[23] | 83% †§ |

(b) CUAVE Video

Table 1: Classification performance for visual speech classification on (a) AVLetters and (b) CUAVE. Learning sparse RBM features improve performance. The deep autoencoders perform the best and show effective cross modality learning. §These results consider continuous speech recognition, although the *normal* portion of CUAVE consists of speakers saying isolated digits. †These models use a visual front-end system that is significantly more complicated than ours and a different train/test split.

On the CUAVE dataset (Table 1b), there is an improvement by learning video features with both video audio compared to learning features with only video data. The deep autoencoder models ultimately performs the best, obtaining a classification score of 69.7%. In our model, we chose to use a very simple front-end that only extracts bounding boxes (without any correction for orientation or perspective changes). A more sophisticated visual front-end in conjunction with our models has the potential to do even better.

The video classification results show that the deep autoencoder model achieves cross modality learning by discovering better video representations when given additional audio data. In particular, even though the AVLetters dataset did not have any audio data, we were able to obtain better performance by learning better video features using other unlabeled data sources which had both audio and video data.

However, we also note that cross modality learning did not help to learn better audio features; since our feature learning mechanism is unsupervised, we find that our model learns features that adapt to the video modality but are not useful for speech classification.

### 4.4 Multimodal Fusion Results

Although using audio information alone performs reasonably well for speech recognition, fusing audio and visual information can substantially improve performance, especially when the audio is degraded with noise [19, 20, 21, 23]. Hence, we evaluate our models in both clean and noisy audio settings.

| Feature Representation | Accuracy (Clean Audio) | Accuracy (Noisy Audio) |
|---|---|---|
| (a) Best Audio-Only | 95.8% | 79.6% |
| (b) Best Video-Only | 69.7% | 69.7% |
| (c) Bimodal Deep Autoencoder | 90.0% | 77.6% |
| (d) Best-Video + Best-Audio | 87.0% | 75.5% |
| (e) **Bimodal + Best-Audio** | **94.4%** | **81.6%** |

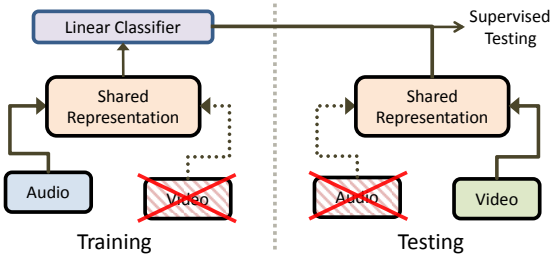
Table 2: Digit classification performance for bimodal speech classification on CUAVE, under clean and noisy conditions. We added white Gaussian noise to the original audio signal at 0db SNR. Best audio refers to the best audio features we learned (single layer RBM for audio). Best video refers to the video-only deep autoencoder features (Table 1b).

The video modality complements the audio modality by providing information such as place of articulation that can help distinguish between similar sounding speech. However, when one simply concatenates audio and visual features (Table 2(d)), it is often the case that performance is worse as compared to using only audio features. Since our models are able to learn multimodal features

that go beyond simply concatenating the audio and visual features, we propose combining the audio features with our multimodal features. When the best audio features are concatenated together with the bimodal features (Table 2(e)), we achieve an increase in accuracy in the noisy setting. This shows that the learned multimodal features are better able to complement the audio features.

### 4.5 Shared Representation Learning

While the above results show that we have learned useful features for video and audio, it does not yet show that the model captures correlations across the modalities. In this experiment, we assess if multimodal features indeed form a *shared* representation that has some invariance to audio or video inputs. During supervised training, we provide the algorithm data solely from one modality (e.g., audio) and later tested only on the other modality (e.g., video). In essence, we are telling the supervised learner how the digits "1", "2", etc. *sound* like and asking it to figure out how to *visually* recognize digits – "hearing to see" (Table 3). If our model indeed learns a shared representation that has some invariance to the presented modality, it will be able to perform this task well.




| Train/Test Setting | Accuracy |
|---|---|
| Audio/Video "Hearing to see" | 29.4% |
| Video/Audio "Seeing to hear" | 27.5% |

Table 3: Shared Representation Learning on CUAVE. The diagram (above) depicts the Audio/Video "Hearing to see" setting.

On the "hearing to see" task, the deep autoencoder obtains an accuracy of 29.4%, while simple baselines perform at chance (10%). Similarly, on the "seeing to hear" task, the model obtains 27.5%. This shows that our learned shared representation is partially invariant to the input modality.

### 4.6 Visualization of learned features

By visualizing our features, we found that the visual bases captured lip motions and articulations. In particular, the learned features include different mouth articulations, opening and closing of the mouth, exposing teeth, among others. We present some visualizations of the learned features in Figure 4.

Figure 4: Visualization of Learned Representations. These figures correspond to two deep hidden units, where we visualize the most strongly connected first layer features. The units are presented in audio-visual pairs (we have found it generally difficult to interpret the connection between the pair).

### 4.7 McGurk effect

The McGurk effect [1] refers to an audio-visual perception phenomenon where a visual */ga/* with a audio */ba/* is perceived as */da/* by most subjects. Since our model learns a multimodal representation, it would be interesting to see if the model was able to replicate a similar effect. We obtained data from 23 volunteers speaking 5 repetitions of */ga/*, */ba/* and */da/*.

Using the learned bimodal deep autoencoder features, we trained a linear SVM on a 3-way classification task. The model was tested on three conditions that simulate the McGurk effect. When the visual and audio data matched at test time, the model was able to predict the correct class

| Audio / Visual Setting | Model prediction | | |
|---|---|---|---|
| | */ga/* | */ba/* | */da/* |
| Visual */ga/*, Audio */ga/* | 82.6% | 2.2% | 15.2% |
| Visual */ba/*, Audio */ba/* | 4.4% | 89.1% | 6.5% |
| Visual */ga/*, Audio */ba/* | 28.3% | 13.0% | 58.7% |

Table 4: McGurk Effect

*/ba/* and */ga/* with an accuracy of 82.6% and 89.1% respectively. On the other hand, when a visual */ga/* with a voiced */ba/* was mixed at test time, the model was most likely to predict */da/*, even though */da/* neither appears in the visual or audio inputs. This is consistent with the McGurk effect on people.

### 4.8 Additional Control Experiments

Recall that we trained the bimodal deep autoencoder with two-thirds of data having one modality missing. To evaluate the role of such a training scheme, we performed a control experiment where we trained the bimodal deep autoencoder without removing any of the modalities. In this experiment, we found that training without any missing data resulted in inferior performance.[6] By inspecting the models, we found that training without missing data led to more modality specific units in the shared representation layer. Conversely, the model trained with the data with missing modalities had more connections to both modalities in the shared representation layer. This supports the hypothesis that having training data with missing modalities is required for the model to learn a shared representation and show cross modality learning.

To evaluate whether a deep architecture is needed or a shallow one would suffice, we trained a bimodal shallow model by training a sparse RBM over the concatenated audio and video data (Figure 2b). However, the correlations between the audio and video modality are highly non-linear and not easily captured by a shallow model. As a result, we find that the model learns largely separate audio and video features. In particular, we find hidden units that have strong connections to variables from either modality but few units that connect across the modalities. Thus, the shallow model is effectively learning two separate representations.

## 5 Related Work

While we present special cases of neural networks here for multimodal learning, we note that prior work on audio-visual speech recognition [13, 24, 25] has also explored the use of neural networks. Yuhas et al. [24] trained a neural network to predict the auditory signal given the visual input. They showed improved performance in a noisy setting when they combined the predicted auditory signal (from the network using visual input) with a noisy auditory signal. Duchnowski et al. [13, 25] trained separate networks to model phonemes and visemes and combined the predictions at a phonetic layer to predict the spoken phoneme. They also attempted combining the representations using the hidden layer from each modality.

In contrast to these approaches, we explicitly use the hidden units to build a new representation of our data. Furthermore, we do not explicitly model phonemes or visemes, which require expensive labeling efforts. Finally, we build deep bimodal representations by modeling the correlations across the learned shallow representations.

## 6 Conclusion

Hand-engineering task-specific features is often difficult and time consuming. For example, it is not immediately clear what the appropriate features should be for lipreading with visual only data. This difficulty is more pronounced with multimodal data as the features have to relate multiple disparate data sources. In this paper, we employed deep learning architectures to learn multimodal features from unlabeled data and also to improve single modality features through cross modality learning.

### Acknowledgments

We thank Clemson University for providing the CUAVE dataset and University of Surrey for providing the AVLetters2 dataset. We also thank Quoc Le, Andrew Saxe, Andrew Maas, and Adam Coates for insightful discussions, and the anonymous reviewers for helpful comments. This work is supported by the DARPA Deep Learning program under contract number FA8650-10-C-7020.

[6] Performance of bimodal deep autoencoder without augmented dataset: Video-only tasks (Table 1) - 50.4% on AVLetters1 and 62.1% on CUAVE. "Hearing to see" and "Seeing to hear" tasks - at chance.

## References


[1] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[2] Q. Summerfield. Lipreading and audio-visual speech perception. *Trans. R. Soc. Lond.*, pages 71–78, 1992.

[3] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. In *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.

[4] R. Raina, A. Battle, H. Lee, and B. Packer. Self-taught learning: Transfer learning from unlabeled data. In *ICML*, pages 759–766, 2007.

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

[6] G. Hinton, S. Osindero, and Y.W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[7] R. Salakhutdinov and G. Hinton. Semantic hashing. *IJAR*, 50(7):969–978, 2009.

[8] H. Lee, C. Ekanadham, and A. Ng. Sparse deep belief net model for visual area V2. In *NIPS*, 2007.

[9] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002.

[10] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[11] P. Vincent, H. Larochelle, Y. Bengio, and P.A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103. ACM, 2008.

[12] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, pages 886–893. IEEE, 2005.

[13] P. Duchnowski, U. Meier, and A. Waibel. See me, hear me: Integrating automatic speech recognition and lip-reading. In *ICSLP*, pages 547–550, 1994.

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

[15] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *ICASSP*, volume 2. Citeseer, 2002.

[16] I. Matthews, T.F. Cootes, J.A. Bangham, and S. Cox. Extraction of visual features for lipreading. *PAMI*, 24, 2002.

[17] S. Cox, R. Harvey, Y. Lan, and J. Newman. The challenge of multispeaker lip-reading. In *International Conference on Auditory-Visual Speech Processing*, pages 179–184, 2008.

[18] W. Fisher, G. Doddington, and Goudie Marshall. The DARPA speech recognition research database: Specification and status. In *DARPA Speech Recognition Workshop*, 1986.

[19] 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.

[20] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Multimodal fusion and learning with uncertain features applied to audiovisual speech recognition. In *MMSP*, 2007.

[21] V. Pitsikalis, A. Katsamanis, G. Papandreou, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation. In *ICSLP*, pages 2458–2461, 2006.

[22] P. Lucey and S. Sridharan. Patch-based representation of visual speech. In *HCSNet Workshop on the Use of Vision in Human-Computer Interaction*, 2006.

[23] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):423–435, 2009.

[24] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 1989.

[25] U. Meier, W. Hürst, and P. Duchnowski. Adaptive Bimodal Sensor Fusion For Automatic Speechreading. In *ICASSP*, pages 833–836, 1996.