AUDIO-VISUAL DEEP LEARNING FOR NOISE ROBUST SPEECH RECOGNITION

Jing Huang and Brian Kingsbury

IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
{jghg,bedk}@us.ibm.com

ABSTRACT

Deep belief networks (DBN) have shown impressive improvements over Gaussian mixture models for automatic speech recognition. In this work we use DBNs for audio-visual speech recognition; in particular, we use deep learning from audio and visual features for noise robust speech recognition. We test two methods for using DBNs in a multimodal setting: a conventional decision fusion method that combines scores from single-modality DBNs, and a novel feature fusion method that operates on *mid-level* features learned by the single-modality DBNs. On a continuously spoken digit recognition task, our experiments show that these methods can reduce word error rate by as much as 21% relative over a baseline multi-stream audio-visual GMM/HMM system.

Index Terms— Audio-visual speech recognition, Deep belief networks, Noise robustness

1. INTRODUCTION

Motivated by the bimodality (auditory and visual) of human speech perception [1] and by the need for robust automatic speech recognition (ASR) in noisy environments, significant research effort has been directed into the study of audio-visual speech recognition (AVSR). Because visual data provides a stream of information that is separate from the audio and invariant to acoustic noise, AVSR has shown noticable improvements over audio-only speech recognition in both clean and noisy conditions [2, 3, 4, 5, 6, 7, 9].

The most successful AVSR systems extract visual features from the facial region of interest and combine them with acoustic features using multi-stream hidden Markov models (HMMs). It has been demonstrated that multi-stream *decision* fusion attains significant improvement in recognition accuracy over single-stream *feature* fusion methods [10]. Discriminative training techniques such as minimum phone error (MPE) [11], MPE trained features (fMPE) [12], and speaker adaptation techniques can be naturally extended to multi-stream HMM based AVSR systems by applying these methods to each stream, either independently or jointly [13, 14]. However, it was found that fMPE does not generalize well and the large gains from fMPE did not carry over to mismatched test conditions.

Deep belief network (DBN) acoustic models have achieved impressive improvements over Gaussian mixture models (GMMs) for automatic speech recognition [15]. One possible reason for this performance difference between DBNs and GMMs is that the distributed representation induced by a DBN is especially well suited to modeling data, such as speech, that is influenced by many different sources of variability. Supporting this notion is recent work showing that in the deeper layers of the network, a DBN acoustic model learns a representation of speech that is less influenced by speaker characteristics than standard features such as MFCCs [16].

Because a neural network makes minimal assumptions about the distribution of the training data, it can potentially generalize better than a GMM, and it easily accommodates multiple input feature streams, simplifying feature fusion. Neural network acoustic models are trained to estimate the posteriors of acoustic classes (HMM states) given the acoustic features, and because they are normalized, acoustic posterior probabilities are more easily used in multi-stream decision fusion systems than acoustic likelihoods. Consequently, neural networks have been a popular tool in audio-visual speech recognition [17, 18, 19, 20]. As a special case of neural networks, DBNs were recently applied to multimodal deep learning [21]. In particular, a DBN was used to learn better features from both modalities and showed its effectiveness in speech classification and visual speech tasks.

In this paper we investigate the use of DBNs to improve audio-visual speech recognition. Because visual-only performance is far worse than audio-only performance, our focus is on how to extract better audio-visual features using a DBN, with the goal being better speech recognition in noisy conditions. In other words, we are interested in using the visual modality to supplement the audio. We investigate two techniques to achieve this. The first technique is a decision fusion technique in which two single-modality DBNs, one for audio and one for visual features, are trained as posterior estimators for HMM states, and the decisions from the single-modality models are combined. The second technique is a novel feature fusion method that combines *mid-level* features learned by the single-modality DBNs. On a continuously spoken digit recognition task, our experiments show that these methods can reduce word error rate by as much as 21% relative over a baseline multi-stream audio-visual GMM/HMM system.

The paper is structured as follows. The baseline multi-stream audio-visual HMM system is reviewed in Section 2. Section 3 briefly describes the DBN training process. The decision fusion and mid-level feature fusion methods are explained in Section 3.2, while the experimental setup is described in Section 4 and results are reported in Section 5. We briefly review some related work in Section 6, and conclusions are drawn in Section 7.

2. THE BASELINE MULTI-STREAM HMM SYSTEM

We briefly describe our baseline multi-stream GMM/HMM AVSR system, including extraction of audio and visual features and the decision fusion process for the audio and visual streams.

To obtain audio features, 24 mel frequency cepstral coefficients (MFCCs) are computed over a sliding window of 25 msec at a frame rate of 100 Hz. Next, the MFCCs are mean-normalized, supervectors are formed by splicing together 9-frame windows of MFCCs, and the supervectors are projected via linear discriminant analysis (LDA) to a 40-d feature space.

Appearance-based visual features are extracted from an automatically estimated region of interest (ROI) surrounding the speaker's mouth by applying a two-dimensional separable DCT to the ROI (please see [9] for the details of ROI extraction), and retaining thetop 100 coefficients with energy. The resulting vectors then go though a pipeline consisting of intra-frame LDA, upsampling to match the audio frame rate, and feature mean normalization, producing a 30-dimensional feature stream ([9]). To model inter-frame dynamics, 15 consecutive frames are spliced and projected to 40 dimensions with another LDA transform.

In the multi-stream HMM decision fusion approach, each stream is modeled by a separate HMM, where the HMMs share the same topology and context-dependent state set. Our baseline uses GMMs to estimate audio and visual class-conditional emission probabilities $P_a(\mathbf{o}_{a,t}\,|c)$ and $P_v(\mathbf{o}_{v,t}\,|c)$, respectively, where $c \in C$ denotes the context-dependent HMM states. The stream likelihoods are combined using [4]

$$P_{av}(\mathbf{o}_{av,t}\,|c) = P_a(\mathbf{o}_{a,t}\,|c)^{\lambda} \times P_v(\mathbf{o}_{v,t}\,|c)^{1-\lambda} \qquad (1)$$

where λ is used to appropriately weight the contribution of each stream. In this work, a fixed value of λ is used, although it may also be time-dependent [22] or adaptive [23].

3. AUDIO-VISUAL DBN AVSR

3.1. Audio and Visual DBNs

The DBNs used in this work go through two training phases: a generative pre-training phase that initializes the weights to a good location in weight space, and a discriminative fine-tuning phase in which the network is trained to perform a specific classification task. The pre-training is done in a greedy, layer-wise fashion [24]. First, the input weight layer is trained, in an unsupervised fashion, as a restricted Boltzmann machine (RBM) using contrastive divergence. Next, the input weights are frozen and the second layer is trained as an RBM on the hidden representations produced by the first layer. This process continues until all hidden layers have been initialized. Because the input features are continuous variables, the first layer is trained as a Gaussian-Bernoulli RBM, while subsequent layers are trained as Bernoulli-Bernoulli RBMs.

Once pretraining is complete, the output layer of weights, which is not generatively pretrained, is initialized with small random values, then the entire network is trained in a supervised fashion using backpropagation with the cross-entropy loss function. For supervised training, the data is randomized at the frame level and organized into mini-batches of 128 frames each. A held-out set of data is used for adjusting the learning rate and determining when training has converged. Specifically, if the held-out loss improves by less than 1% after a complete pass over the training data, the learning rate is reduced by a factor of two. After the learning rate has been reduced five times, training ends. Note that both the audio and visual DBNs have exactly the same structure and go through the same training procedure: they differ only in their input features.

The trained DBNs estimate posterior probabilities, $P(c|\mathbf{o}_t)$. For Viterbi decoding, these posteriors are normalized by the class priors, $P(c)$, to convert them to scaled likelihoods.

3.2. Learning a combined audio-visual representation

The simplest way to learn a combined representation from audio and visual features is to concatenate them at the input to a DBN. While this approach jointly models the distribution of both modalities, it is limited in that it will be difficult for units to learn cross-modal correspondences when both modalities are influenced by many different sources of variability, such as lighting conditions and speaker characteristics [21]. Previous AVSR work confirms that a simple feature fusion approach is inferior to decision fusion [7, 10]. As an alternative, we investigate the fusion of mid-level features learned by modality-specific DBNs, as illustrated in Figure 1. The expectation is that the mid-level features will be less influenced by extraneous sources of variability, making it easier to learn cross-modal correlations, and potentially outperforming simple decision fusion on the single-modality DBNs. We test two different ways of using the audio-visual representation: either the combined audio-visual DBN is used directly as an acoustic models (with the posteriors converted to scaled likelihoods, as described above), or we compute probabilistic features from the audio-visual DBN and use them as features for a conventional GMM acoustic model. The probabilistic features are computed by projecting the 100-dimensional input to the softmax nonlinearity to 40 dimensions using an LDA transform computed from audio alignment of the training data.

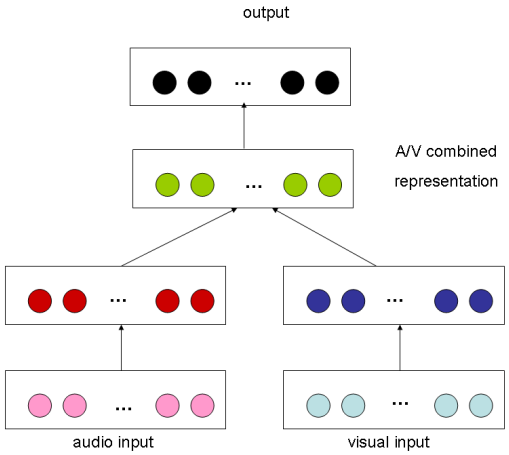

Fig. 1. *DBN trained from A/V combined representation*

4. EXPERIMENTAL SETUP

Our experiments are conducted on a continuous digit recognition audio-visual database collected with the IBM infrared headset [9]. In contrast to most of audio-visual data collection with uniform studio lighting, our data were collected in different office lighting conditions. The data set consists of a total of 107 subjects uttering approximately 35 connected sequences of 7 or 10 digits. We split the 107 speakers into training and testing sets: 70 speakers (about 3.8 hours) are used for training, and the remaining 37 speakers (about 1.5 hours) are used for testing. There are no overlapping speakers in the training and the testing data. Both training and testing sets have an average SNR of 20dB. In addition to the clean test data, which matches the training data, another noisy test set is built by artificially corrupting the test set with additive "speech babble" noise, resulting in an average SNR of 7dB. Recognition results are presented on both clean and mismatched noisy test sets.

The baseline GMM/HMM model uses three-state, left-to-right phonetic HMMs with 100 quinphone context dependent states

and 2100 diagonal-covariance Gaussian mixture components. The stream combination uses fixed weights of 0.7 on the audio stream and 0.3 on the video stream in all test conditions.

The DBNs classify their inputs into the same set of 100 context-dependent HMM states used by the baseline and use the same audio and visual feature processing. In order to have a reasonable comparison between the GMM/HMM baseline and the DBN systems, the DBNs are designed so that the total number of trainable parameters is roughly comparable to the baseline GMMs. We tested two input sizes for the audio and visual DBNs, with one model set taking only a single frame of input, and another taking a 3-frame context. All DBNs have five hidden layers with logistic nonlinearities and a softmax nonlinearity at the output. In the 1-frame DBN the first two hidden layers contain 256 units each, while the remaining three contain 128 units. Therefore the total number of parameters in the 1-frame DBN is about the same as that of the baseline HMM. All hidden layers in the 3-frame DBN contain 256 units; therefore, the total number of parameters in 3-frame DBN is about twice the size of the baseline HMM. We use 5% of the training data (3 speakers) as held-out data.

5. RESULTS

Results are presented as word error rate (WER) for audio-only (A), visual-only (V) and decision-fusion of audio-visual (AV) recognition. Table 1 compares the baseline HMM models and various DBNs on the matched clean test data and mismatched noisy test data. First we notice that the audio DBN performance lags behind the baseline on the matched test condition. This is due to the small amount of training data and held-out data. The tiny gains from 3-frame-input DBNs also confirms that the amount of training data is an issue. However, the focus of this paper is on noise robust speech recognition.

Audio-visual DBNs outperform the baseline GMM/HMM on the noisy condition, improving WER from 13.4% to 12.4% (7% relative improvement), while using a comparable number of parameters. While audio DBN performance is behind the baseline on the matched test, the synergy between AV DBNs are much better than AV HMMs: the gain from AV DBN decoding is 36% relative to audio performance, and the gain from AV HMM decoding is only 24% relative to audio performance.

System	Match			Noisy		
	A	V	AV	A	V	AV
baseline	1.7	35.2	1.3	26.5	35.2	13.4
1-frame	2.2	35.7	1.4	24.3	35.7	**12.4**
3-frame	2.1	35.7	1.4	24.2	35.7	12.2

Table 1. *Comparison of baseline model and DBN models on matched clean and mismatched noisy test data. Both the baseline and DBN models use decision fusion.*

Next, we present results for a mid-level feature fusion method that concatenates hidden representations from the audio DBN and visual DBN and uses the result as input to a third, audio-visual DBN. We test configurations that use the second (L2), third (L3), or fourth (L4) layer representations from the single-modality, 1-frame input DBNs. The AV DBN has three hidden layers with logistic nonlinearities and a softmax nonlinearity at the output. For the L2 input, the AV DBN has 512 inputs and 800 units in the first hidden layer, while for the L3 and L4 inputs it has 256 inputs and 400 units in the first hidden layer. In all cases the second and third hidden layers contain 128 units, and the AV DBNs have the same 100 context-dependent HMM state output targets. We also investigate the use of probabilistic features computed from the audio-visual DBNs, where the features are computed by projecting the 100-dimensional input to the softmax nonlinearity to 40 dimensions using an LDA transform, and then these features are used with a standard GMM acoustic model. Note that this approach involves the use of many more trainable parameters than either the baseline GMM/HMM AVSR or the DBN/HMM AVSR that uses decision fusion.

Table 2 compares the L2, L3, and L4 audio-visual DBNs, and corresponding GMM/HMM models trained with probabilistic DBN features. On the matched test, the performance degrades with the audio-visual combined representation. This is consistent with results from feature-fusion HMM-based AVSR and with results from [21], where the concatenation of audio-visual features performed worse than audio alone in matched test conditions. However, in the noisy case, the combined representation shows improvement over audio-visual DBN decoding, from 12.4% to 11.7%. With the probabilistic DBN features, the gain is much more, from 12.4% to 10.6% from the L3 DBN. Compared to the baseline of 13.4%, the relative gain is 21%. Because the LDA projection used to compute the probabilistic features is based on audio-only alignments, it helps to choose the right audio-visual feature set for speech recognition. Even though the gains on the matched condition are small from the projection, it helps significantly on the noisy test. For the L2 DBN features, it improves from 15.5% to 11.3%, and for the L3 DBN features, it improves from 13.0% to 10.6%. This is exactly what we are hoping for.

Input	Match	Noisy
	WER	WER
L2	4.0	15.5
L2-projected	3.6	11.3
L3	2.7	13.0
L3-projected	2.4	**10.6**
L4	2.8	11.7
L4-projected	2.3	11.5

Table 2. *Comparison of different combined A/V representation models on matched clean and mismatched noisy test data.*

6. RELATION TO PRIOR WORK

In [18] 2-layer NNs were trained for phonemes and visemes. The paper compared three combination approaches: combination on phonetic layer (decision fusion), combination at the input layer (feature fusion) and combination at the hidden layer. The experimental task is speaker-dependent continous spelling of German letter strings (8 letters on average), with 170 sequences from one speaker for training and 30 test sequences from the same speaker at different noise levels. The results showed the combination at the phonetic layer was the best with adaptive weighting schemes. Our study also examines the combination of audio and visual representations from different layers, but our focus is on noise robustness. Also, we work on a larger, speaker-independent task.

Lewis and Powers presented some preliminary research on using psycholinguistic knowledge and showed that late integration (i.e. de-

cision fusion) was better than early integration (i.e. feature fusion) in AVSR with neural networks. While this knowledge was well known in AVSR, the authors showed in NN framework different ways of integrating audio-visual features. The experiments were on classification of 9 phonemes, 3 visemes and 3 voicing groupings. Again training data is rather small, collected from 3 subjects with 2 examples of each phoneme/position pair (positioned at 1.5 or 1.8 meters away from recoding device).

The idea of learning a combined audio-visual representation from the hidden representations of audio and visual DBNs is inspired by [21], which focused on cross-modality feature learning for visual speech classification. The experiments were interesting and combined diverse datasets of CUAVE, AVLetters, AVLetters2, Stanford Dataset and TIMIT. The results showed that a video deep autoencoder achieved cross-modal learning, obtaining better visual representations when given additional audio input. However, the audio deep autoencoder did not enjoy similar improvements: adding visual input could hurt performance. Our results on the matched test condition in this paper confirm this point. Our work differs in that it uses a bimodal DBN to find a noise-robust audio-visual speech representation, rather than a better representation for lipreading.

7. CONCLUSIONS AND DISCUSSIONS

In this paper we have shown that AVSR with DBNs performs better than AVSR with GMM/HMM models on a mismatched noisy condition, and that a GMM/HMM using combined audio-visual DBN features can outperform an audio-visual DBN/HMM system. On a continuously spoken digit recognition task, our experiments show that a GMM/HMM trained from the combined audio-visual representation reduces WER by 21% relative over audio-visual multi-stream GMM/HMM models on mismatched noisy data. In order for DBNs to be successfully applied, more training data and large vocabulary collection would be necessary.

8. REFERENCES

[1] W.H. Sumby and I. Pollack (1954), "Visual contribution to speech intelligibility in noise," in *J. Acoustical Society America*, 26: 212–215.

[2] T. Chen and R.R. Rao (1998), "Audio-visual integration in multi-modal communication," in *Proc. IEEE*, 86(5): 837–852.

[3] A. Janin, D. Ellis and N. Morgan (1999) "Multi-stream speech recognition: Ready for prime time?", in *Proc. Europ. Conf. Speech Technol.*, pp. 591–594, 1999.

[4] S. Dupont and J. Luettin (2000), "Audio-visual speech modeling for continuous speech recognition," in *IEEE Trans. Multimedia*, 2(3): 141–151.

[5] C.C. Chibelushi, F. Deravi, and J.S.D. Mason (2002), "A review of speech-based bimodal recognition," in *IEEE Trans. Multimedia*, 4(1): 23–37.

[6] M. Heckmann, F. Berthommier, and K. Kroschel (2002), "Noise adaptive stream weighting in audio-visual speech recognition," in *EURASIP J. Appl. Signal Process.*, 2002(11): 1260–1273.

[7] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior (2003), "Recent advances in the automatic recognition of audio-visual speech," in *Proc. IEEE*, 91(9): 1306-1326.

[8] G. Potamianos (2006), "Audio-Visual Speech Recognition," in *Encyclopedia of Language and Linguistics*, Second Edition, (Speech Technology Section - Computer Understanding of Speech), K. Brown (Ed. In Chief), Elsevier, Oxford, United Kingdom, ISBN: 0-08-044299-4, 2006.

[9] J. Huang, G. Potamianos, J. Connell and C. Neti (2004), "Audio-visual speech recognition using an infrared headset," in *Speech Communication* 44(4), 83–96.

[10] E. Marcheret, S. Chu, V. Goel, G. Potamianos (2004), "Efficient Likelihood Computation in Multi-Stream HMM Based Audio-Visual Speech Recognition," in *Int. Conf. Speech and Language Processing*, 2004.

[11] D. Povey and P. C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," in *Proceedings of ICASSP*, 2002.

[12] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proceedings of ICASSP*, 2005.

[13] J. Huang and D. Povey, "Discriminatively Trained Features Using fMPE for Multi-Stream Audio-Visual Speech Recognition," in *Proceedings of Interspeech*, 2005.

[14] J. Huang and K. Visweswariah, "Combined Discriminative Training for Multi-Stream HMM-based Audio-Visual Speech Recognition," in *Proceedings of Interspeech*, 2009.

[15] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," in *IEEE Signal Processing Magazine*, 29(6): 82–97, 2012.

[16] A. Mohamed, G. Hinton, G. Penn, "Understanding how Deep Belief Networks perform acoustic modelling," in *Proceedings of ICASSP*, 2012.

[17] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lipreading," in *Proceedings of ICSLP*, 1994.

[18] U. Meier, W. Hurst and P. Duchnowski, "Adaptive Bimodal Sensor Fursion for Automatic Speechreading," in *Proceedings of ICASSP*, 1996.

[19] T. Lewis and D. Powers, "Audio-Visual Speech Recognition using Red Exclusion and Neural Networks," in *Journal of Research and Practice in Information Technology*, 2003.

[20] M. Kim, J. Ryu, and E. Kim, "Speech Recognition by Integrating Audio, Visual and Contextual Features Based on Neural Networks," in *Advances in Natural Computation*, Lecture Notes in Computer Science, 2005.

[21] J. Ngiam, A. Khosla, J. Nam, H. Lee and A.Ng, "Multimodal Deep Learning", in *International Conference on Machine Learning*, 2011.

[22] A. Garg, G. Potamianos, C. Neti, T. Huang, "Frame-Dependent Multi-Stream Reliability Indicators for Audio-Visual Speech Recognition," in *Int. Conf. Acoustic Speech and Signal Processing*, 2003.

[23] V. Pitsikalis, A. Katsamanis, G. Papandreou, and P. Maragos. "Adaptive multimodal fusion by uncertainty compensation," in *Proceedings of ICSLP*, 2006.

[24] G. E. Hinton, S. Osindero, and Y. Teh. "A Fast Learning Algorithm for Deep Belief Nets," in *Neural Computation*, vol. 18, pp. 1527-1554, 2006.