Open-Domain Audio-Visual Speech Recognition: A Deep Learning Approach

Yajie Miao, Florian Metze

Carnegie Mellon University
{ymiao,fmetze}@cs.cmu.edu

Abstract

Automatic speech recognition (ASR) on video data naturally has access to two modalities: audio and video. In previous work, audio-visual ASR, which leverages visual features to help ASR, has been explored on restricted domains of videos. This paper aims to extend this idea to open-domain videos, for example videos uploaded to YouTube. We achieve this by adopting a unified deep learning approach. First, for the visual features, we propose to apply segment- (utterance-) level features, instead of highly restrictive frame-level features. These visual features are extracted using deep learning architectures which have been pre-trained on computer vision tasks, e.g., object recognition and scene labeling. Second, the visual features are incorporated into ASR under deep learning based acoustic modeling. In addition to simple feature concatenation, we also apply an adaptive training framework to incorporate visual features in a more flexible way. On a challenging video transcribing task, audio-visual ASR using our proposed approach gets notable improvements in terms of word error rates (WERs), compared to ASR merely using speech features.

Index Terms: Automatic speech recognition, audio-visual ASR, deep learning

1. Introduction

The pervasive deployment of speech interfaces requires automatic speech recognition (ASR) systems to handle various types of variability. In recent years, the introduction of deep neural networks (DNNs) has achieved the state-of-the-art recognition accuracy on a wide range of acoustic modeling tasks [1, 2, 3]. In general, DNNs display superior recognition accuracy than the traditional Gaussian mixture models (GMMs) [4]. However, robustness remains to be a challenge for DNN models [5]. For example, in [6], it is revealed that the performance of DNNs degrades significantly as the signal-to-noise ratio (SNR) drops. An effective strategy to deal with variability is to incorporate additional knowledge explicitly into DNN models. On this aspect, [7, 8, 9, 10, 11] study the incorporation of i-vectors as speaker representations to smooth out the effect of speaker variability. To handle the variability of the distance between speakers and microphones, [12] proposes to learn a DNN-based extractor to model the speaker-microphone distance information dynamically on the frame level. Then distance-aware DNNs are built by appending these descriptors to the acoustic features as the DNN inputs.

Another line of work has focused on improving robustness of acoustic models with *audio-visual ASR*. The process of speech perception is bi-modal in nature. This motivates researchers to take advantage of visual features to enhance ASR, especially under noisy conditions [13, 14, 15, 16]. However, these previous proposals have limitations in that they generally adopt visual features extracted from the speaker's mouth region, including lip contours and mouth shapes. Although available in highly constrained videos, these features are not always obtainable from open-domain videos. For example, in a large portion of YouTube videos, the speakers do not appear in the video frames at all. Another limitation of traditional audio-visual ASR is the requirement for alignments between speech and video frames. Since the speech and video streams have different sampling rates, aligning them may introduce inaccurate visual features into acoustic modeling.

In this paper, we aim to relax these constraints and extend audio-visual ASR to open-domain videos. Our approach to achieving this is based on deep learning, and can be split into two parts:

- We extract the visual features using deep architectures, i.e., deep convolutional neural networks (CNNs). Depending on the types of visual information we attempt to model, the deep CNN networks are trained on object recognition or scene labeling tasks. This enables us to obtain informative visual features directly from the raw pixels of open-domain videos. Also, with the CNN models, we can generate visual features on the segment level (rather than the frame level), which removes the need for time synchronization between the audio and video streams.
- We investigate audio-visual ASR in the context of DNN-based acoustic models. In addition to simple feature concatenation, we also apply an adaptive training framework to incorporate visual features in a more flexible way. Together with our previous work [10, 11, 12], we prove the generality of this adaptive training approach in fusing different types of additional information into DNN acoustic models.

Our experiments are conducted on a task of transcribing open-domain amateur videos. Compared with the baseline DNN, our audio-visual DNN model achieves significant improvement in terms of word error rates (WERs). More experiments are presented to examine the utility of different types of visual features in enhancing ASR.

2. Related Work on Audio-Visual ASR

The bimodal nature of speech perception motivates researchers to work on audio-visual ASR. The goal of this field is to explore the visual modality to improve the performance of ASR, especially under noisy acoustic conditions with low SNRs. There exist two key problems for audio-visual ASR. The first problem is the extraction of the visual features that can potentially benefit ASR. Previous work [13, 14, 15, 16] has generally exploited visual features that are extracted from the speaker's mouth region. For example, in the automatic lip-reading literature [17, 18], areas of interest (AOI) centering around the lip are extracted

to form the image features. In [13], coefficients of lip shape and intensity, together with their temporal dynamics, are generated as the visual descriptors. A more straightforward feature type is the gray-scale pixel values of the (downsampled) image covering the speaker's mouth [19]. In [16], visual features are obtained from pixel color using raster scan, i.e., 30-dim RGB features with 10 dimensions for each color. The second problem lies in the fusion of the audio and visual modalities into the bimodal system. In practice, this fusion can be conducted either on the feature level [14, 20, 16] where a bimodal front-end is constructed with the two feature streams, or on the decision level [21, 13] where outputs from classifiers are combined during the recognition stage. A major challenge for this fusion is the asynchrony of the audio and video streams. To solve this issue, attempts have been made to combine the classifier outputs (e.g., states likelihoods) at a coarser level, for example the phoneme or even the word level [21, 13]. Another useful tool to deal with this asynchrony is dynamic Bayesian networks (DBNs) which allow for different levels of information fusion and have shown effectiveness in audio-visual ASR [22, 23].

Despite these advancement, audio-visual ASR still has limitations that prevent its deployment to real-world video data. For example, mouth/lip related features are not always available in open-domain videos. In this paper, we aim to further remove these constraints and thus achieve truly open-domain audio-visual ASR.

3. Audio-Visual ASR using Deep Learning

This section presents our framework to perform audio-visual ASR on open-domain videos. This framework consists of two major steps: extracting visual features from the video stream and incorporating visual features into acoustic models. Both steps are conducted via deep learning.

3.1. Extraction of Visual Features

Our previous work [24] has investigated speaker attributes in helping acoustic modeling. For a particular video, we may find frames (images) where the speaker shows up. Face recognition enables us to obtain attributes (e.g., age, gender, etc.) of the speaker from these frames. However, the application of speaker attributes is limited in that these features cannot be extracted accurately from videos where the speakers make no appearances at all. To further mitigate this limitation, we propose to extract visual features that are extracted directly from raw images.

Suppose we are dealing with an utterance u which has the acoustic features $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$, where T is the total number of speech frames. On a video transcribing task, there always exists a video segment corresponding to u. This video segment is represented as $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N\}$, where N is the number of video frames. The video frames are sampled normally at a lower sampling rate than the speech frames, i.e., $N < T$. From this video segment, we select a video frame $\mathbf{v}_n$ which serves as the image representation for the speech utterance. Then two types of image features are extracted from $\mathbf{v}_n$.

Object Features. Our first type of visual information is derived from object recognition, the task on which deep learning has accomplished tremendous success [25]. The intuition is that object features encapsulate information regarding the acoustic environment/condition of speech data. For example, classifying an image to the classes "computer keyboard" and "monitor" indicates that the speech segment has been recorded in an office. We extract this object information using a deep CNN model which has been trained on a comprehensive object recognition dataset, e.g., ImageNet [26], and the resulting CNN model is referred to as *object-CNN*. Then, on our target ASR task, the video frame $\mathbf{v}_n$ is fed into the CNN model, from which we get the distribution (posterior probabilities) over the object classes. These probabilities encode the object-related information that are finally incorporated into DNN acoustic models. Due to the high dimension of the probabilities, we may need to perform dimension reduction. More details regarding training of the CNN networks will be presented in Section 4.1.

Place Features. The utility of the object features comes from the "place" information that is implicitly encoded by the object classification results. It is then natural to utilize place features in a more explicit way. To achieve this, we train a deep CNN model meant for the scene labeling task. Given a video frame, the classification outputs from this *place-CNN* encode the place information, which is then incorporated into acoustic models. For convenience of formulation, the resulting visual feature vector for this utterance u is represented as $\mathbf{f}_u$.

3.2. Incorporation of Visual Features

After obtaining the visual features, we adopt two methods to incorporate the visual features into DNN acoustic models.

3.2.1. Feature Concatenation

A simple way is to append the visual features to the original DNN inputs. The DNN model is then built over the augmented feature vectors. Within the utterance u, the augmented feature vector on the t-th frame now becomes $[\mathbf{o}_t, \mathbf{f}_u]$. During fine-tuning, the bottom layers in the DNN are trained to fuse the visual information and the acoustic features with non-linear transformations. The activations from these bottom layers become more robust to environment variability, and thus benefit phonetic states classification performed by the upper layers.

3.2.2. Visual Adaptive Training

In our previous work [10, 11], we have presented a framework to perform speaker adaptive training (SAT) for DNN models. This approach requires an i-vector [27] to be extracted for each speaker. Based on the well-trained speaker-independent (SI) DNN, a separate *adaptation neural network* is learned to convert i-vectors into speaker-specific linear feature shifts. Adding these shifts to the original DNN inputs produces a speaker-normalized feature space. Parameters of the SI-DNN are re-updated in this new space, which finally generates the SAT-DNN model. This framework has also been applied successfully to descriptors of speaker-microphone distance, and thus achieves *distance adaptive training* (DAT) for DNNs [12].

In this paper, we port this idea to the visual features, which enables us to conduct *visual adaptive training* (VAT) for DNNs. Specifically, in the SAT framework, we replace the i-vector representation with the visual features. An adaptation network is learned to take the visual features as inputs and generate an adaptive feature space with respect to the visual descriptors. Note that in this case, the linear feature shifts generated by the adaptation network are utterance-specific rather than speaker-specific. Re-updating the parameters of the DNN in the normalized feature space gives us the adaptively trained VAT-DNN model. This VAT-DNN model takes advantage of the visual features as additional knowledge, and generalizes better to unseen variability.

4. Experiments

4.1. Experimental Setup

4.1.1. Dataset

Following [12, 24], our acoustic modeling task is to transcribe real-world *instructional videos*. To create the dataset, we download a collection of English videos from online video archives. These videos are uploaded by social media users to share expertise on specific tasks (e.g., oil change, sandwich making, etc.). ASR on these videos is challenging because they have been recorded in various environments (e.g., office, kitchen, baseball field, train, etc.). After data preparation, we finally get 94 hours of speech, out of which 90 hours are selected for training and 4 hours for testing. For decoding, a trigram language model (LM) is trained on the training transcripts. This LM is then interpolated with another trigram LM trained on an additional set of 300 hours of instructional-video transcripts.

4.1.2. Object-CNN

The object-CNN network for object feature extraction follows the standard AlexNet architecture [25]. The network contains 5 convolution layers which use the rectifier non-linearity (ReLU) [28] as the activation function. In the first and second convolution layers, a local response normalization (LRN) layer is added after the ReLU activation, and a max pooling layer follows the LRN layer. In the third and fourth convolution layers, we do not apply the LRN and pooling layers. In the fifth convolution layer, we only apply the max pooling layer, without LRN applied. 3 fully-connected (FC) layers are placed on top of the convolution layers. The first and second FC layers have 4096 neurons, whereas the number of neurons in the last FC layer is equal to the number of classes.

This object-CNN is trained on ImageNet [26], a dataset containing over 15 million labeled images belonging to around 22,000 categories. We use a subset of ImageNet created for the 2012 Large-Scale Visual Recognition Challenge (ILSVRC). The ILSVRC training set amounts to 1.2 million images covering 1000 classes. Each image from the training data is resized to 256x256, and then normalized with mean and variance normalization. From this resized image, we crop the 227x227 block from the center as the inputs into the CNN. Therefore, the CNN inputs have the size of 3x227x227, where 3 is the number of channels (RGB). Model training optimizes the standard cross-entropy (CE) objective. The resulting object-CNN achieves a 20% top-5 error rate on the ILSVRC 2012 testing set.

4.1.3. Place-CNN

In order to extract place information, we train the place-CNN network on the MIT Places dataset [29] which contains 2.5 million images belonging to 205 scene categories. Examples of the scenes include "dinning room", "coast", "conference center", "courtyard", etc. We use the complete set of 2.5 million images for training, and follow the same image pre-processing as used on ImageNet (Section 4.1.2). The architecture of the place-CNN is almost the same as that of the object-CNN. The only difference is that in the final FC layer, the place-CNN has 205 neurons corresponding to the 205 scene classes, whereas the object-CNN contains 1000 neurons.

4.1.4. Visual Features

The trained object-CNN or place-CNN can be transferred to our video transcribing task. Each speech utterance corresponds to a video segment, from which we *randomly* select an image frame. The same pre-processing steps as described in Section 4.1.2 are applied to the image. Feeding this image to the CNN models gives us the classification results over the classes, either objects or places. The dimension of the posterior probabilities is reduced down to 100[1] via principal component analysis (PCA). The PCA transform is estimated only on the training set of the video transcribing dataset. The resulting 100-dim feature vectors are taken as the additional visual information.

4.2. Results of Object Features

Table 1 shows the results of the DNN models when object features are incorporated. When we perform VAT, the 100-dim utterance-level visual features are taken as the inputs into the adaptation network, which contains 3 hidden layers with 512 neurons per layer. From Table 1, we can see that incorporating object features results in moderate but consistent improvement. In particular, when applying the VAT approach, we obtain 0.9% absolute improvement over the baseline DNN (22.5% vs 23.4%), which translates to 3.8% relative improvement.

Table 1: *Results (% WER) of DNN models when object and place features are incorporated. Two approaches are adopted for the incorporation: feature concatenation and VAT.*

Models	Visual Features	WER(%)
DNN (baseline)	—	23.4
Feature Concatenation	100-dim object features	23.0
VAT	100-dim object features	22.5
Feature Concatenation	100-dim place features	22.7
VAT	100-dim place features	22.5

4.3. Results of Place Features

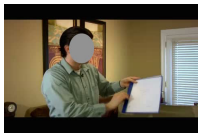
Table 1 shows the results of the final DNN models with place features incorporated. With feature concatenation, the place features turn out to give us the WER 22.7%, a number that is better than the result of feature concatenation with the object features. This reveals that the major benefit of using the visual features is the scene/place information encoded by the image representation. Switching to place-CNN enables us to eliminate the information generated by object-CNN that is unrelated to scenes/places.

The 100-dim place features are further applied in the VAT framework. By doing this, we continue to get gains, bringing the WER down to 22.5%. The VAT model with the place features gets the same WER as the VAT model with the object features. This means that although containing noise, the object features can be transformed into a more effective representation with the adaptation network embedded in the VAT framework.

4.4. Results of Further Combination

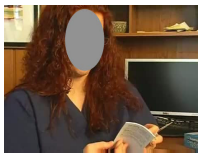
We further employ the adaptive training approach to integrate more types of additional information, and show the results in Table 2. First of all, the place features are combined with speaker-related visual features, including the 11-dim speaker attributes described in [24], as well as some 50-dim speaker action features. To obtain the action information, we feed each

[1]The choice of 100 is intended to match the dimension of i-vectors used in previous work [7, 10, 11]. This consistency facilitates experimental comparisons.

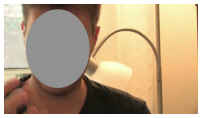


(a) Indoor, WER 27.55% → 27.55%

(b) Indoor, WER 22.27% → 22.27%



(c) Indoor, WER 16.42% → 16.42%



(d) Indoor, WER 26.44% → 26.44%

Figure 1: Examples of videos on which incorporating place information gives no improvement. For each video, we show an image frame extracted from it. "A% → B%" means that on this video, the WER is improved from A% to B%.

(a) Kitchen, WER 30.95% → 27.38%

(b) Baseball field, WER 18.70% → 15.65%

(c) Airport apron, WER 34.12% → 28.24%

(d) Train, WER 44.71% → 38.24%

Figure 2: Examples of videos on which incorporating place information gives over 10% relative improvement.

of the video segments into an action recognition system. This system has been trained on the UCF101 dataset [30] that consists of realistic action videos coming from 101 categories (e.g., baby crawling, surfing, sky diving). Action recognition gives us a 101-dim vector containing the probabilities over the 101 classes. The feature dimension is further reduced to 50 through PCA. Appending the speaker attributes and actions to the place features results in 161-dim enlarged visual features. Adaptive training corresponding to these features produces the WER of 22.3%, which is slightly better than adaptive training purely with the place features (22.5%).

Finally, we combine the visual features with speaker i-vectors. Following previous work [7, 10, 11], we extract a 100-dim i-vector for each speaker, and combine the i-vector with the 161-dim visual features. With this more comprehensive feature representation, we are able to bring the WER down to 21.5% using adaptive training, i.e., 8.1% relative improvement compared to the original DNN baseline.

Table 2: *Results (% WER) of the adaptive training approach when different types of additional information are employed.*

Visual Features	WER(%)
161-dim combined visual features	22.3
100-dim speaker i-vectors	22.0
261-dim i-vectors + visual features	21.5

4.5. Qualitative Analysis

In addition to the overall recognition accuracy, we also want to analyze the effects of incorporating the place information on individual videos. Specifically, for each video, we compare the WER achieved by the feature concatenation model with the WER achieved by the baseline DNN model. Figure 1 exemplifies the videos on which feature concatenation ends up to have the same WER as the baseline DNN model, i.e., incorporating the place features results in no improvement. In our training data, a large majority of the videos have been recorded in indoor environments (e.g., offices). This data distribution enables the indoor environments to be sufficiently modeled by the acoustic model. That is why in Figure 1, adding the additional place information generates no gains on indoor videos.

In comparison, Figure 2 exemplifies the videos on which feature concatenation outperforms the baseline DNN model by at least 10% relatively. We observe that most of these videos are recorded either in outdoor environments (e.g., baseball field, airport apron, street, etc.) , or in non-typical indoor conditions (e.g., kitchen, music studio, etc.) where music/noise severely interferes with the actual speech. Adding the place descriptors helps the DNN model normalize the acoustic characteristics of these rare conditions, and thus benefits the generalization to unseen testing speech.

5. Conclusions and Future Work

In this paper, we have attempted to extend audio-visual ASR to open-domain videos. A unified deep learning framework is presented to achieve this. We extract utterance-level visual features using deep learning architectures which have been pre-trained on object recognition or scene labeling tasks. The visual features are then incorporated into DNN acoustic models, via feature concatenation or adaptive training. For our future work, we would like to study the adaptation of language models using the visual features. The visual features studied in this paper are potentially useful for language modeling. For example, objects recognized from the video stream correlate with words (presumably nouns) appearing in the transcripts. A straightforward approach would be to train RNNs based language models and attach the visual features to the RNN inputs. Also, we are highly interested to incorporate the visual features into more advanced acoustic models, e.g., Long Short-Term Memory (LSTM) models [31, 32] and end-to-end ASR pipelines [33].

6. Acknowledgements

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. This work was partially funded by Facebook, Inc. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Facebook, Inc.

7. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[3] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[4] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[5] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 4, pp. 745–777, 2014.

[6] Y. Huang, D. Yu, C. Liu, and Y. Gong, "A comparative analytic study on the Gaussian mixture and context dependent deep neural network hidden Markov models," in *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014.

[7] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 55–59.

[8] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6334–6338.

[9] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 225–229.

[10] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014.

[11] ——, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 11, pp. 1938–1949, 2015.

[12] Y. Miao and F. Metze, "Distance-aware dnns for robust speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2015.

[13] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *Multimedia, IEEE Transactions on*, vol. 2, no. 3, pp. 141–151, 2000.

[14] T. Chen, "Audiovisual speech processing," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 9–21, 2001.

[15] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 1–6.

[16] Y. Kashiwagi, M. Suzuki, N. Minematsu, and K. Hirose, "Audio-visual feature integration based on piecewise linear transformation for noise robust automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 149–152.

[17] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 1. IEEE, 1993, pp. 557–560.

[18] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: integrating automatic speech recognition and lip-reading." in *ICSLP*, vol. 94, 1994, pp. 547–550.

[19] B. P. Yuhas, M. H. Goldstein Jr, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *Communications Magazine, IEEE*, vol. 27, no. 11, pp. 65–71, 1989.

[20] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual lvcsr," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 165–168.

[21] M. J. Tomlinson, M. J. Russell, and N. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE, 1996, pp. 821–824.

[22] J. N. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "Dbn based multi-stream models for audio-visual speech recognition," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004.

[23] V. Estellers and J.-P. Thiran, "Overcoming asynchrony in audio-visual speech recognition," in *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*. IEEE, 2010, pp. 466–471.

[24] Y. Miao, L. Jiang, H. Zhang, and F. Metze, "Improvements to speaker adaptive training of deep neural networks," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[27] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[28] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, vol. 15, 2011, pp. 315–323.

[29] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[30] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[31] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014.

[32] Y. Miao and F. Metze, "On speaker adaptation of long short-term memory recurrent neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2015.

[33] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015.