Are You Speaking: Real-Time Speech Activity Detection via Landmark Pooling Network

Boyu Wang[1] and Xiaolong Wang*[2]
[1] Computer Science Department, Stony Brook University, NY, USA
[2] IBM, CA, USA
boywang@cs.stonybrook.edu, xiaolong.wang@ibm.com

***Abstract*— In this paper, we propose a novel visual information based framework to solve the real-time speech activity detection problem. Unlike conventional methods which commonly use the audio signal as input, our approach incorporates facial information into a deep neural network for feature learning. Instead of using the whole input image, we further develop a novel end-to-end landmark pooling network to act as an attention-guide scheme to help the deep neural network only focus the related portion of the input image. This helps the network to precisely and efficiently learn highly discriminative features for speech activities. What's more, we implement a recurrent neural network with the gated recurrent unit scheme to make use of the sequential information from video to produce the final decision. To give a comprehensive evaluation of the proposed method, we collect a large-scale dataset from unconstrained speech activities, which consists of a large number of speech/non-speech video sequences under various kinds of degradation. Experimental results demonstrate the superiority of our proposed pipeline over previous approach in terms of performance and efficiency.**

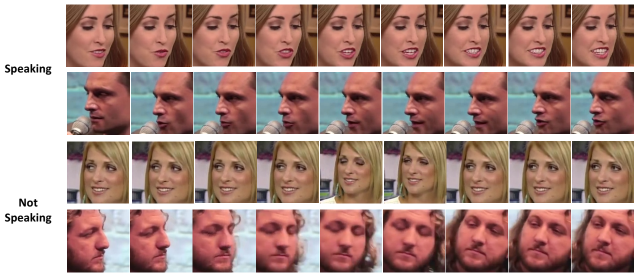

Fig. 1: **Typical video sequences in the LSW dataset.** The top two rows show two speech sequences. The bottom two rows show two non-speech sequences. Our data covers a wide range of pose variations, e.g., the sequence from the second row shows the person is speaking with half face invisible, and the sequence from the last row shows the person moves his head around but his lip doesn't move.

I. INTRODUCTION

Speech Activity Detection (SAD) remains an essential component in speech processing systems and is an active research area. It is one of the most popular techniques used in speech processing systems such as speech coding and speech recognition. The goal of SAD is to predict whether the user is speaking or not based on the input signal. This technique has various applications such as facilitating speech processing and helping deactivate the processes during the non-speech section, helping save computation and network bandwidth by avoiding unnecessary coding/transmission of silence voice packets through the Internet. A typical SAD system generally consists of three steps. Firstly, the input signal, which can be an audio signal or a visual signal, is collected and pre-processed either by audio recording or images capturing techniques. Secondly, discriminative features are extracted from the input signal to form high-level representations for further processing. Lastly, classification systems are implemented to predict the input signal to be speech or non-speech.

Most of conventional SAD systems rely on audio signals [3, 11, 12, 15]. This is due to the nature that audio is the most straightforward signal to judge whether a person is speaking or not. However, audio signal would easily be influenced by surroundings such as background noise or unstable audio recording. Under such circumstances, the performance of such systems is significantly degraded. To solve this problem, an alternative solution is to use a visual-based SAD pipeline. Instead of using audio signal, this kind of systems mainly rely on the features extracted from visual information [2, 6, 8, 9, 11].

In this paper, we propose a novel visual information based deep neural network framework to solve the SAD problem. Following the nature of SAD systems, which tend to have high requirements of latency and efficiency, we propose a novel Landmark Pooling Network (LPN). Unlike traditional neural networks which usually work on the full image or full feature map, the LPN is able to use facial geometry information to focus only on the small portion of points which are supposed to carry more useful information for the current task. Above this, we implement the recurrent neural network with the gated recurrent unit to model the temporal information from the video sequence to have a better understanding of the input signal. In seek of being in real-time speed, the proposed model is designed in an end-to-end manner and to be comparatively small compared to other deep learning frameworks.

Our work has two main contributions. Firstly, we propose one novel LPN scheme, which is able to use only a small portion of the input image to achieve the speech detection functionality precisely and effectively. Secondly, we collect a new large-scale dataset from unconstrained

* indicates the corresponding author.

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

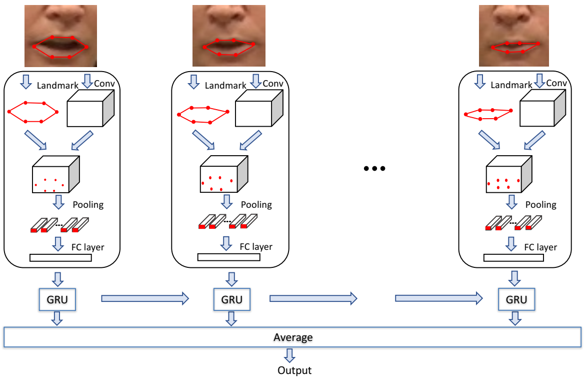

Fig. 2: **Network Architecture using landmark pooling feature learning.** The network starts with one convolutional layer, followed by a landmark pooling layer and a fully connected layer. Then, features are fed into the Recurrent Neural Network (RNN) Gated Recurrent Unit (GRU) for aggregation. A fully connected layer is applied at every time step, followed by an average operation. A softmax layer is added at the end for classification. We only visualize 6 landmarks for demonstration.

speech activities. To our best knowledge, this is the first well-labeled unconstrained visual-based SAD dataset with good data samples. The rest of this paper is organized as follows: Section II describes the details of our new dataset. Section III explains the proposed architecture. Section IV reports the performance of the model, as well as experimental results analysis and comparisons. Section V draws the conclusion of our work and briefly discusses the future work.

II. LSW Dataset description

Pose variation has long been an issue in visual-based SAD as well as unconstrained face detection/recognition problems. To our best knowledge, none of previous work or datasets on SAD have taken the pose variation issue into consideration. However, this issue is nontrivial since large pose variations may lead to significant face feature shift, which may further lead to detection/recognition failures. To address the SAD problem in the wild, we build a new large-scale dataset: Labeled Speech in the Wild (LSW).

The LSW dataset consists of a large number of speech and non-speech video sequences, which cover a wild range of face pose variations and viewpoint variations. All the video sequences are collected from YouTube and are about “panel discussion” or “roundtable discussion”, since videos under such theme settings are more likely to involve multiple speakers with large face pose variations. In total, we collect 45 different YouTube videos. Several typical video sequences from our LSW dataset are shown in Fig. 1. The top and bottom two rows represent speech sequences and non-speech sequences, respectively. The second row gives an example where the person is speaking while the face is half invisible. The last row illustrates the example where the person moves his head around while the lip doesn’t move. Compared to other existing datasets on this problem, the LSW dataset is larger, more challenging, and closer to real-life scenarios.

Set	Num. of subjects	Num. of sequences
Train	171	8002
Test	24	901
Total	195	8903

TABLE I: Statistics of LSW dataset.

In total, we have collected 5512 speech and 3391 non-speech sequences. There are 195 subjects in total. Most sequences have the length of 40 frames. We split them into a training set and a test set. The split is across different videos, there are no overlaps between the training set and the test set. We list the statistics of our dataset in Table I. The dataset is public available at `http://vision.cs.stonybrook.edu/~boyu/LSW_dataset.zip`

III. Network Architecture

In this section, we introduce the whole pipeline of the proposed SAD system. The goal of our system is to determine whether one video sequence as speech or non-speech. We treat the detection problem as a binary classification problem. Our network takes both the raw images and facial landmarks as input. Fig. 2 illustrates our whole network architecture. The pipeline mainly consists of two parts: the LPN for fast and accurate high-level feature learning and the RNN for temporal information modeling and classification.

Our system takes a video sequence as input. At each time step, the current frame is fed into a convolutional layer (with 64 filters of size 7×7 and stride 2), followed by a landmark pooling layer, which uses landmark locations to pool convolutional feature maps. We will describe the details of LPN in subsection III-A. The LPN can help the network effectively focus on a certain region of interest, so as to pay more attention to the changes in the mouth outlines. We use 20 landmarks around the mouth region for LPN and concatenate the pooled features into a 20×64-dimension vector. Then we apply one fully-connected layer to project this vector into a 64-dimension high-level feature vector. Furthermore, at each time step, the extracted high-level features are fed into the Recurrent Neural Network (RNN) with Gated Recurrent Unit (GRU) for aggregation. We use one layer of GRU cell with RNN size of 64 (dimension of the memory vectors). The output hidden states of RNN at every time step will be averaged together. Then a fully connected layer is added to map the former extracted features into a two-dimensional vector. A softmax layer is added at the end for speech and non-speech classification. We use leaky rectified linear unit (Leaky-ReLU) [13] as our activation function. The details of each building block in our network are given in the following subsections.

A. The Landmark Pooling Network

When dealing with the SAD problems, it is natural to believe that the mouth region is the most important part of the face. Motivated by this fact, we design a convolutional neural network (CNN) using only the specific landmark points around the mouth region. These landmarks perform as an attention-guide mechanism for the network.

Here we describe the details of our LPN. Given a convolutional feature map F of size $H \times W \times C$, where H, W, C are the height, width, and depth of the feature map respectively (in our case, $H = 20$, $W = 20$, $C = 64$), and a 2D facial landmark location vector v of size $L \times 2$, where L is the number of landmarks to use, in our case, $L = 20$, the procedure to learn the landmark-pooling features are as follows:

- For each landmark located at (x_i, y_i) on the input image, we map it to the corresponding locations $(\hat{x_i}, \hat{y_i})$ on the feature map F, such that $0 \leq \hat{x_i} < W$, $0 \leq \hat{y_i} < H$.
- Select the feature vector on F via a landmark location $(\hat{x_i}, \hat{y_i})$, which is computed by $f_i = F(\hat{x_i}, \hat{y_i}, :)$.
- Concatenate all f_i together to produce f as the pooled feature.

The top part of Fig. 2 illustrates the whole process. It can be trained using standard backpropagation: the gradients only propagate through the landmark locations.

The LPN benefits our system in several ways: 1). Fewer parameters. Compared to working on the whole feature map, landmark pooling can significantly reduce the number of parameters to learn, so as to reduce the output feature map size. 2) A better attention mechanism. The network will focus more on the outline of the mouth, which helps the network assign higher weights to important locations. Another point worth mentioning is that instead of locating the landmarks on the original input image, we do it on the feature map after the convolutional layer. This is due to the fact that each pixel on the feature map has a receptive field of 7×7 on the original image. By doing so, the features are learned from the landmark's all neighbor pixels, which is more reliable than only using the landmark pixel. We use the Recurrent Neural Network [14] with Gated Recurrent Unit (GRU) to model the temporal information of previous extracted spatial features at every time step. GRU [4] is a recently proposed variant of RNN, which is based on Long Short-Term Memory (LSTM) architecture [10] but with a simpler form. We adopt GRU to the proposed pipeline due to its simpler architecture and good performance on sequence modeling problems.

B. Implementation Details

The network is trained in an end-to-end manner. The cross-entropy loss is used as the loss function and the backpropagation through time (BPTT) algorithm is applied for optimization. We use Adagrad [7] as the optimizer with the learning rate initialized by 0.0001 and then reduced by a factor of 10 after every 50k steps. And the learning was stopped after 200k iterations. The network parameters are initialized by a Gaussian distribution with zero mean and standard deviation of 0.01. We use a momentum coefficient of 0.9 and a weight decay factor of 0.0005. The whole system is implemented via Tensorflow [1].

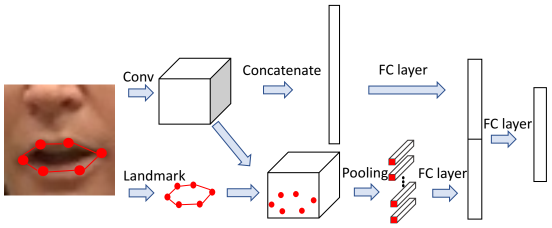

Fig. 3: **LPN + Appearance CNN**

IV. Experiments

A. Data preprocessing

To pre-process the input data, for each face image, we crop out the mouth region based on the landmarks' locations. Then we convert the cropped image to grayscale and resize it to 40×40. The data is normalized to zero mean and unit variance. In order to improve robustness and avoid overfitting, data augmentation is applied. First, we augment data by applying the same transformation to images in a single sequence, e.g. random flipping, random cropping, random distortion. Second, face movement speed is varied by random deletion or duplication of images, and by randomly changing the frame rate of the sequence (from 0.8 to 1.2 times of original frame rate). Third, more non-speech sequences are generated by randomly choosing one frame from speaking sequence and repeating this frame to a certain length.

B. Baselines

We compare the proposed network with three standard baselines.

Landmark Only: This baseline only uses the landmark information. Following [5], we take the Fourier transform for the landmark sequence, which can separate the lip movements to different frequency bins. A two-layer perceptron with leaky-ReLU activation is trained on this frequency spectrum to make decisions. We refer this method as *LO* for short.

Appearance CNN: This baseline only uses the appearance information and takes the whole image as input. The network consists of one convolutional layer, followed by a spatial max pooling layer. One fully-connected layer is used at the end for feature extraction. After the features are extracted at each time step, they are fed into a standard GRU for feature aggregation. We refer this method as *ACNN* for short.

Landmark + Appearance CNN: This baseline uses both appearance and geometry information to make a prediction. The network has two branches, one for appearance, and the other one for the landmark. For the appearance sub-branch, it is the same as *ACNN*. For the geometry sub-branch, the features are extracted using two fully-connected layers. Subsequently, the feature vector from the appearance branch and geometry branch are concatenated together, and then fed into GRU for feature aggregation.

Method	Accuracy
Landmark Only	66.2
Appearance CNN	76.7
Landmark + Appearance CNN	77.2
LPN	72.1
LPN + Appearance CNN	**79.9**

TABLE II: Performance comparisons among all the baselines methods and the proposed methods on the LSW dataset.

C. Evaluation

Baselines: We report the classification accuracies of all methods on the test set. The LO method results in the lowest accuracy 66.2%, which is reasonable due to the shallow architecture and large pose variations which result in poor landmark detection. The ACNN method uses more facial information and a deeper network architecture, which improves the accuracy to 76.7%. By simply fusing the appearance and the landmark features together with the CNN architecture (L+ACNN), we can improve the accuracy to 77.2%. Even though the appearance features already include the landmark features, we can see that by emphasizing the landmark representation, we can further improve the overall performance. This proves that the landmark features is more valuable in this task.

Our results: We achieved 72.1% accuracy by using the proposed LPN method, which is 6% higher than the original LO method, thanks to the deep network structure, but still 5% lower than the ACNN network. This is reasonable since LPN only uses 5% facial information compared to ACNN. Finally, by combining LPN with ACNN (LPN+ACNN), we achieved the best performance at 79.9% on this dataset, which indicates that our method can better integrate geometry cues with landmark information. The architecture is shown in Fig. 3. The performance comparisons are illustrated in Table II.

D. Online Deployment for Active Speech Detection

We deploy the model for real-time speech detection. We adopt a sliding window technique. At each time step, we look back up to a fixed length of previous T frames (which is 20 in our case), and evaluate this acquired sequence with our model. The result indicates whether the sequence contains any speech activities or not from T frames back to now.

Table III shows the number of parameter comparisons and speed comparisons among different methods on both GPU and desktop. For GPU, we use a GTX 1080Ti. For desktop, we use MacBook Pro with Intel core i5 and 8G RAM memory. In the evaluation, we assume that all mouth images and landmarks are already available, which is a fair assumption, as the landmarks can be detected in real time. As shown in Table III, compared to ACNN, our LPN method has 10 times less parameters and is around 30% faster than ACNN on both GPU and CPU in speed.

E. Case study

Fig. 4 illustrates some failure cases. Generally, for a speech sequence, it's difficult for our system to make an accurate decision if the mouth region is rarely moving.

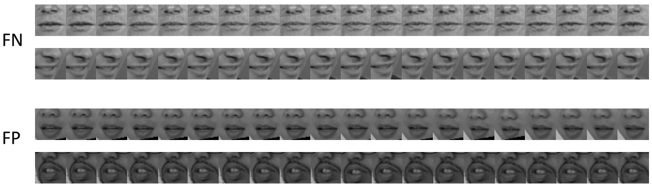

Fig. 4: **Failure Cases.** The top two rows are false negative (FN) sequences, where the ground truth is labeled as speaking, but our method predicts as non-speaking. The bottom two rows are false positive (FP) sequences, where the ground truth is labeled as non-speaking, but our method predicts as speaking. These failure cases are even confusing for human.

Method	#params	GPU (fps)	Desktop(fps)
LO	28,609	1658	4133
ACNN	1,211,841	152	109
L+ACNN	1,248,641	187	128
LPN	110,017	198	138

TABLE III: Speed evaluation for different methods

The top two rows are false negative (FN) cases, where the ground truth is labeled as speaking, but our method predicts as non-speaking. Taking a closer look at the video sequence, we can notice that even though the person is speaking, the mouth region rarely changes. This is challenging to our system since we heavily rely on the correct visual information. Using the audio signal to reinforce our system can be a good solution to solve the problem.

The bottom rows are false positive (FP) cases, where the ground truth is labeled as non-speaking, but our method predicts as speaking. The sequence in the third row is also confusing for human, the mouth gradually closes up, so our method predicts it as speaking. In the fourth row, the person's mouth is widely open but doesn't move. This is probably due to the fact that there are lots of sequences with wide-open mouths labeled as speaking sequences in our dataset. This indicates that for further work, we should build a model that can better utilize temporal information.

V. Conclusion

In summary, we studied the visual-based Speech Activity Detection problem in this paper. We proposed a novel landmark pooling network scheme, which guides the network to focus on small portion regions so as to work more precisely and more efficiently. We utilized the recurrent neural network with the gated recurrent unit for temporal information modeling. We also design an efficient system with comparable fast detection speed. Our method is evaluated on a new large-scale challenging dataset collected from unconstrained speech activities under various kinds of degradation. Experiments on multiple evaluation settings show that our model is both accurate and fast. We will explore more on capturing the temporal information part to improve the performance.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[2] P. Chakravarty, S. Mirzaei, T. Tuytelaars, et al. Who's speaking?: Audio-supervised classification of active speakers in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 87–90. ACM, 2015.

[3] P. Chakravarty, J. Zegers, T. Tuytelaars, et al. Active speaker detection with audio-visual co-training. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 312–316. ACM, 2016.

[4] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[5] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

[6] R. Cutler and L. Davis. Look who's talking: Speaker detection using video and audio correlation. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 3, pages 1589–1592. IEEE, 2000.

[7] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[8] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 27(5):545–559, 2009.

[9] F. Haider and S. Al Moubayed. Towards speaker detection using lips movements for humanmachine multiparty dialogue. *FONETIK 2012*, page 117, 2012.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] J. Kheradiya, S. Reddy, and R. Hegde. Active speaker detection using audio-visual sensor array. In *Signal Processing and Information Technology (ISSPIT), 2014 IEEE International Symposium on*, pages 000480–000484. IEEE, 2014.

[12] D. Li, C. Taskiran, N. Dimitrova, W. Wang, M. Li, and I. Sethi. Cross-modal analysis of audio-visual programs for speaker detection. In *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, pages 1–4. IEEE, 2005.

[13] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.

[14] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010.

[15] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.