INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Real Time Online Visual End Point Detection Using Unidirectional LSTM

Tanay Sharma, Rohith Aralikatti, Dilip Kumar Margam, Abhinav Thanda
Sharad Roy, Pujitha A. K, Shankar M Venkatesan

Samsung R&D Institute India, Bangalore

{tanay.sharma, r.aralikatti, dilip.margam, abhinav.t89}@samsung.com,
{sharad.roy, pujitha.k, s.venkatesan}@samsung.com

Abstract

Visual Voice Activity Detection (V-VAD) involves the detection of speech activity of a speaker using visual features. The V-VAD is useful in detecting the end point of an utterance under noisy acoustic conditions or for maintaining speaker privacy. In this paper, we propose a speaker independent, real-time solution for V-VAD. The focus is on real-time aspect and accuracy as such algorithms will play a key role in detecting end point especially while interacting with speech assistants. We propose two novel methods one using CNN and the other using 2D-DCT features. Unidirectional LSTMs are used in both the methods to make it online and learn temporal dependence. The methods are tested on two publicly available datasets. Additionally the methods are also tested on a locally collected dataset which further validates our hypothesis. Additionally it has been observed through experiments that both the approaches generalize to unseen speakers. It has been shown that our best approach gives substantial improvement over earlier methods done on the same dataset.

Index Terms: V-VAD, Visual End Point Detection, Unidirectional LSTMs, CNN, DCT

1. Introduction

For the last two decades considerable research has been done on the use of multiple modalities (specifically audio and visual) in improving various speech technologies such as speech recognition[1, 2, 3, 4, 5, 6, 7], speech enhancement [8, 9, 10], speech separation [11, 12] , voice activity detection [13, 14, 15, 16] all of which establish the importance of using visual modality in speech.

End point in speech refers to the time when user stops speaking. With the advent of speech assistants and commercial robots, hands free execution of speech commands has become an indispensable feature for speech recognition systems. Therefore, the need for accurately detecting end point has increased many folds. An incorrect or delayed detection can adversely affect the computational efficiency and accuracy of the speech recognition system.

To mitigate this problem an accurate and efficient end point detection module is required. The end point detection or more generally voice activity detection has been studied quite extensively in the previous years and it has been observed that more emphasis is given to audio signal as they are more informative than visual signal. But from the past decade many audio visual and only visual voice activity detection approaches [13, 14, 15, 16] have also been studied. The reason being that in a noisy environment audio loses its information and thus visual modality can be used to increase the overall performance of voice activity detection as visual modality is independent of noise. To detect that ambient surrounding is noisy, people have used SNR estimation approaches such as [17].

Moreover, recently the field of lip reading has grown in importance[1, 5, 6, 7]. Thus accurately detecting end point using visual features has become indispensable for such applications. End point has to be detected online i.e., the decision whether an end point is reached has to be made after each incoming video frame. Therefore, it has be to be highly efficient both in terms of accuracy and latency.

In addition, V-VAD is expected to function even on unseen speakers. In general V-VAD accuracies tend to be higher for speaker dependent/seen speaker cases but the accuracy drops on unseen speakers [18, 15]. Unseen speakers along with unseen sentences i.e., speakers and spoken sentences both are not in training set is yet another challenge for V-VAD. This is due to the fact that unseen sentences can have relatively different temporal dependencies which is not seen while training the dataset.

Our work discusses two methods for V-VAD which differ in the visual features used. The first method uses features from Convolutional Neural Network(CNN) extracted from lip images of the speaker. The second method uses vectorized 2D-DCT features obtained from lip images. Recent works [19, 16, 13, 14] on audio as well as visual VAD, applied Long Short Term Memory (LSTM) in learning long term dependencies. Both of our methods also incorporate LSTMs. Specifically, we used unidirectional LSTM as bidirectional LSTM will require the whole sequence to be fed at a time which defeats the purpose of making the whole system online. Unidirectional LSTM gives us the liberty to feed each frame separately.

Our experiments on different datasets in a variety of conditions show the robustness of our methods. The results are compared with similar works done in the past. The key contributions of this paper can be summarized as follows:

- A light weight efficient model to detect end point using visual features when audio features cannot be used especially in a very noisy place or for privacy concerns.
- Comparison with existing methods.
- Experiments in speaker independent scenario and on unseen sentences to show the robustness of the system.

The remainder of our paper is organized as follows. Section 2 describes the related work that has been done using visual features. Section 3 describes the two methods and model architectures. Section 4 presents experiments and results along with information about the datasets that we have used while in Section 5 conclusion and future work are presented.

2. Related Work

Several approaches of VADs including audio, visual and audio-visual features have been explored.

http://dx.doi.org/10.21437/Interspeech.2019-3253

HMM has been a popular choice to model such sequence problems. [20] used separate GMM's for audio and visual and then used multi-stream HMM's to combine both as a model based feature fusion method. In [21] two methods have been proposed for V-VAD. One classifies active appearance parameters of lip region using HMM and other uses retina filtering to understand lip motion. [22] modeled HMM with optical flow features. Real Time AV-VAD is proposed in [23] using geometric features around lip region such as lip contour height followed by 7 point average smoothing and classification of speech/non speech intervals using calculated threshold. Geometric features were also used to analyze the coherency between lip movements and audio for VAD [24]. V-VAD using temporal orofacial features have been explored in [25]. PCA is used to combine different features which goes to EM algorithm to define two clusters of speech and non-speech regions. The similar approach was further extended by improving boundary detection using Bayesian Information Criterion [26]. In [27] the authors used spatio-temporal Gabor transform to measure facial movements in video sequences. V-VAD performance is determined at different speeds using these measurements.

DCT features due to their computational efficiency and better feature representation were used in [18] [28]. MFCC features for audio input and 2D DCT features of the lip region [18] were used for V-VAD and a weighting term was used at decision fusion level. [28] explored the usefulness of profile view along with frontal pose for VAD task using DCT of lip region and GMM modeling.

Recently, CNN features have proved to be efficient in representing the visual data and have been applied in various domains such as large scale image classification [29] and identification of scene in videos [30]. They are also applied in audio domain as features [31] and shows improvements on DNN systems for their generalizing ability and better local correlation of inputs. In V-VAD [15], the authors used CNN features along with early and late fusion for the inclusion of temporal information. It shows high accuracy for the speaker dependent case on grid data [32] but it tends to decrease in speaker independent scenario as temporal information is not captured fully.

Recurrent neural networks (RNN) have proved to be very effective to formulate temporal dependency of sequences. It has been shown in [33] that RNNs outperform GMM HMM model for A-VAD. They used 13 dimensional PLP features and a multi-layer RNN module. Recently, LSTMs which are very special kind of RNN have further shown to outperform simple RNN in learning long term dependencies in data which is useful for most of the sequence learning problems. It has been successfully used in VAD in [19] [16] [13] [14].

An AV-VAD [13] is trained with 26-d handcrafted features as visual features and MFCC features as audio features fed to Bidirectional LSTM layers(BLSTM) separately and their outputs are combined and fed to another set of BLSTM layers and a softmax layer at the end. It has shown to outperform earlier methods. In [14] similar features were used but a modified LSTM that takes care of longer temporal dependencies by including multiple connections to frames in the past. [34] exploited non deep models by using local shape and motion information appearing at spatiotemporal locations of interest for facial video segment and a SVM classifier at the end on unconstrained movie data. [16] proposed a end-to-end framework for A-VAD that combines convolutional neural network (CNN) and RNN. CNN takes raw speech as input extracting relevant features that are fed into the RNN, to get a silence/speech class label per frame. They achieved 4.2% false alarm rates (FAR) under noisy conditions after fixing the system at 2% false reject rate(FRR). Since this approach is explored in speech domain, it will fail in high noise conditions where audio is not informative. Thus we propose a V-VAD using the power of both CNN to model features directly from the raw lip image and LSTM to learn temporal dependencies. We also propose a DCT LSTM approach and compare both the methods. To the best of our knowledge this has not been done for V-VAD.

3. Visual end point detection

3.1. Lip detection and stabilization

The first step in V-VAD is to detect the region of interest (ROI) which in our case is the region surrounding speaker's lips. To detect ROI, we use YOLO v2 [35] model. The YOLO model is trained on RGB lip-images with single class label and outputs the coordinates of the bounding box surrounding the lip. However, even in cases where the subject in the frame is relatively stable, the bounding box may jitter for consecutive frames. To stabilize the bounding box, we use momentum method on the center of the box. The ROI is cropped using the bounding box and converted to grayscale. Grayscale images result in better latency. The ROI is resized to 100x50. We observed empirically that resizing does not affect the V-VAD accuracy but keeps the computational load to a minimum.

3.2. Model architecture

As discussed before we propose two approaches as shown in Fig. 1. In the first approach we use CNN architecture in combination with LSTM layers. We use three 2D CNN layers. Filter size of each CNN layer is 16, 32 and 8 respectively while stride is (2, 2) and kernel size is (5, 5) for each layer. Each CNN layer is followed by a max pooling with a stride and pool size of (2, 2) and a batchnorm layer. ReLU activation is used for non-linearity.

The CNN features at each timestep are fed to two LSTM layers one at a time along with the previous state vector that gives us the updated state after every timestep. Each LSTM layer has 64 hidden units. This state is then passed to a dense layer followed by a softmax layer at the end to determine the speech activity at that instant i.e., the network classifies a particular frame as speech or non-speech. Subsequently end point of the utterance can be inferred using the current and past speech activity results.

In the second approach, we extract 2D-DCT features from ROI and take 100 higher energy components and convert it to a 100-dimensional vector. The DCT features are fed to the same LSTM architecture described above. In other words, the CNN features in the first approach are replaced by DCT features.

3.3. End point detection

The ROI extracted from speaker's video as shown in 3.1 is passed to the model as described in 3.2 which finally outputs speech/no speech labels for each frame. The output labels are passed through a smoothing pipeline which takes an average over a sliding window of 14 frames. This removes inconsistent fluctuations in labels and results in more accurate end point detection. Finally, the smoothed labels are passed to end point detector pipeline. To detect end point we take a window of 21 frames in the past and if 80% or more frames are labelled as silent frames, we consider it as end point. The window size of 21 was selected keeping in mind the usability of the EPD mod-

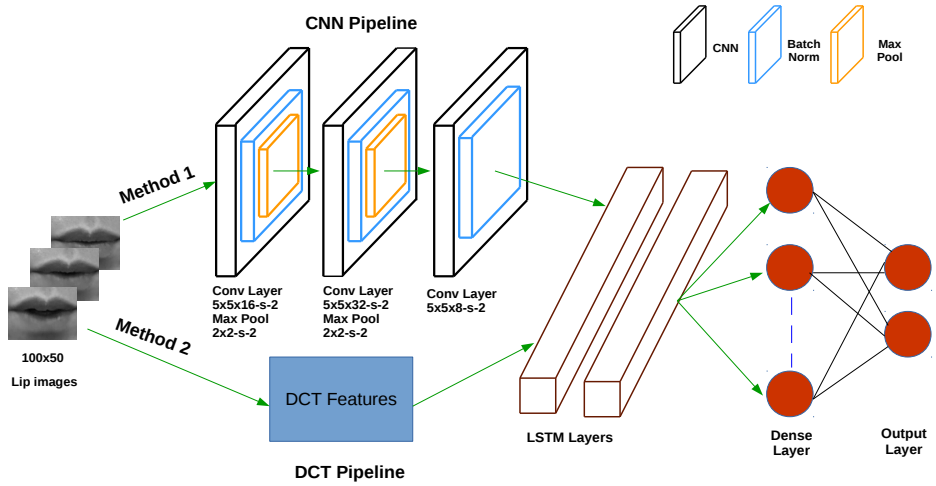

Figure 1: *Model Architecture Diagram*

ule in practical applications. Ideally, we expect that not more than 700ms (21 frames at 30fps) should be taken to detect end point after a user has stopped speaking. If the end-point detection is delayed, the number of silence frames at the end of the sentence increases which affects the accuracy of lip-reading or ASR system. In addition, longer utterances increase the recognition latency. After 21 frames, we continue to estimate end point, but penalize the end point detection accuracy if they are detected after 21 frames till 40 frames. The penalty is varied linearly from 0 to 1 increasing linearly with frame number. After 40 frames, we give 0 weight to the accuracy as mentioned in 1 and 2. The usability of end point detection module will be practically infeasible if it is detected after 40 frames.

$$f_i(n) = \begin{cases} 1 & 0 \leq n \leq 21 \\ 1 - (n-21)/19 & 21 \leq n \leq 40 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where:
n = number of frames after which end point was detected after user stopped speaking
$f_i(n)$ = accuracy of i^{th} test sequence

$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} f_i(n) \quad (2)$$

where:
N = total number of test sequences

4. Experiments and Results

4.1. Dataset

4.1.1. GRID audio-visual corpus

There are many publicly available audio visual datasets. But to train a supervised deep model for V-VAD task, we require a large scale timestamp annotated data. Thus we used a publicly available large audio visual English dataset called GRID [32]. The dataset contains audio visual data of 34 speakers speaking 1000 sentences each along with alignment files for each utterance having word boundaries.From these alignment files, we can easily get speech/no-speech annotations. It is found using the alignments that 200ms-300ms is the average silence period between subsequent words (7-10 frames @30fps). So the model should not confuse silences between words as endpoint.

Therefore, at least 16 frames are required at the end as silent frames to decide the end point. So we filtered out grid data comprising of at least 16 frames at the end as silent frames for training and testing. For speaker dependent experiments on GRID, we use 32 speakers for training, 4 speakers for validation and 3 speakers (300 sentences each) for test. The speakers chosen for validation and test are subset of training speakers. For speaker independent case , we used 27 speakers for training, 3 speakers for validation and 2 speakers for test.

GRID dataset is collected in a controlled environment where all speakers are still with little or no head movement. All videos are recorded in frontal pose. Moreover vocabulary of grid is restricted to 51 words and sentence structure follows the grammer as <command><color><preposition><letter><digit><adverb>. This produces sentences like "Set blue to W 9 soon" which is not practical for everyday usage. So we also created a challenging database to measure our model performance on more practical scenarios.

4.1.2. Indian-English dataset

Our dataset contains 81 Indian speakers speaking 170 English sentences each and is termed as Indian-English Dataset. The sentences are 5-10 words long. Also speakers are asked to hold mobile in hands unlike a fixed camera setup which is not practical in real life usage. We did not put any restrictions apart from the fact that lip should be visible in each frame while speaking. Since mobile was hand held, we were able to capture head motions and different angles similar to the practical scenarios in which people will be communicating with speech assistants. We recorded videos at 1920x1080 resolution. Phoneme alignments were generated using audio files and highly optimized acoustic model. Using the video framerate and aligned timestamps, we were able to get frame wise speech/no speech boundaries for each utterance. The speaker independent experiments on this data was done using train, validation, test split of 70, 5 and 6 speakers respectively. While in speaker-dependent scenario, we

Table 1: *Accuracy of our models on grid where SdA is Speaker dependent accuracy, SiA is Speaker independent accuracy*

Model Name	Grid			
	SdA		SiA	
	Frame Level %	End Point %	Frame Level %	End Point %
DCT2008 [18]	97.66	-	74.68	-
CNN2015 [15]	96.7	-	72	-
DCT LSTM	95	89	91.2	92
CNN LSTM	96.5	99	92.2	97

used 81 speakers for training , 15 speakers for validation and 12 speakers (60 sentences each) for test.

4.1.3. VidTIMIT dataset

We also used another publicly available dataset called VidTIMIT [36] to compare our model performance on sentences that are not in training set. VidTIMIT consists of recordings of 43 people speaking 10 sentences each. The resolution of videos in this dataset is 512x384 which is lower than our dataset. The model was tested on 409 VidTIMIT sentences.

4.2. Results and Discussion

We evaluated our models on the above datasets in different scenarios. Table 1 shows accuracy in speaker dependent and speaker independent scenarios on GRID dataset. Both frame level as well as end point detection accuracy have been listed. For comparison we used two earlier approaches for V-VAD, [15] as CNN2015 and [18] called as DCT2008. Although for speaker dependent case all the approaches seem to perform well but as shown in the table, the accuracy drops drastically for both the earlier approaches in speaker independent case. On the other hand our methods are able to generalize better to unseen speakers. We observe an absolute improvement of 20% from CNN2015 and around 17% from DCT2008 approach. Also the resulting end point accuracy is 97% in speaker independent case which shows efficiency of our model. Since end point detection accuracy was not mentioned in earlier approaches [15, 18] , we compared our results with the available results ie. frame level accuracy. Better frame level accuracy generally follows better end point detection accuracy.

We also conducted similar experiments on Indian English dataset which is an unconstrained and more difficult data in terms of variability of texture, head movements and pose changes. Table 2 shows the results on this dataset.As observed, the frame level accuracy for speaker independent case is 86% which is still around 14% better than CNN2015 and 12% better than DCT2008.

To further validate the generalizability of our method we trained a model by combining GRID and Indian English datasets. This time we kept VidTIMIT dataset as test data to see how our model performs on unseen speakers as well as unseen sentences. Table 3 shows the results. Since this is a very tough scenario as model has not seen such sentences while training, it is expected to get lower accuracies in this experiment. But even in this scenario the CNN method is able to shows 73% frame-level accuracy. Lower accuracies in this case can also be attributed to the low resolution of videos in this dataset. It learns speech movements and doesnot not over-fit to trained sentences. It further validates our hypothesis and shows the robustness of

Table 2: *Accuracy of our models on Indian English dataset*

Model Name	Indian English			
	SdA		SiA	
	Frame Level %	End Point %	Frame Level %	End Point %
DCT LSTM	86	87	85.3	85
CNN LSTM	89.5	90	86.6	88

Table 3: *Accuracy of our model on VidTIMIT Dataset*

Train Data	Test Data	Frame Level Accuracy (%)	
		CNN LSTM	DCT LSTM
Indian English + GRID	VidTIMIT	73	63

system.

To get a measure of how quickly our algorithm identifies end point after a user stops speaking, we have plotted a graph describing percentage of test data in which EPD was detected vs number of frames after user stops speaking in a step size of 5 as shown in Figure 2. The faster it detects after user stops speaking the better it is. As observed in Figure 2, for 90% of the data end point was detected in the given time frame (0-40 frames) while only for 10% of data it was detected before or after the limit.

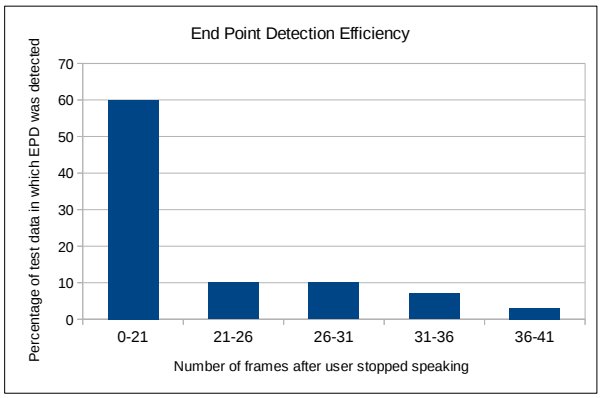

Figure 2: *EPD efficiency*

5. Conclusion

We proposed two methods DCT LSTM and CNN LSTM for end point detection using V-VAD. We showed that learning features directly from raw lip image using CNN seems to outperform DCT. Our approaches seem to outperform other methods in speaker independent scenarios with high margin and thus it can be clearly seen that our model generalizes better to unseen speakers. Further experiments on unseen speaker and unseen sentences showed that these models are not limited to particular domain of training. The small network size and unidirectional LSTMs make it an ideal fit for real time online application. In future work, we plan to make our methods more robust to illumination, extreme poses and missing frames. We are also experimenting with ivectors in visual domain to train a personalized V-VAD for better accuracy for a particular speaker.

6. References

[1] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[2] M. Dilip Kumar, A. Rohith, S. Tanay, T. Abhinav, A. K. Pujitha, R. Sharad, and V. Shankar M, "Lipreading with 3d-2d-cnn blstm-hmm and word-ctc models," *arXiv preprint arXiv:1906:12170*, 2019.

[3] A. Thanda and S. M. Venkatesan, "Multi-task learning of deep neural networks for audio visual automatic speech recognition," 2017.

[4] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 11, 2018. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2018.2889052

[5] K. Xu, D. Li, N. Cassimatis, and X. Wang, "Lcanet: End-to-end lipreading with cascaded attention-ctc," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 548–555.

[6] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: End-to-end sentence-level lipreading," 2016.

[7] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, M. Mulville, B. Coppin, B. Laurie, A. Senior, and N. de Freitas, "Large-scale visual speech recognition," 2018.

[8] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[9] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," *arXiv preprint arXiv:1711.08789*, 2017.

[10] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[11] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[12] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Speaker separation and enhancement using visually-derived speech," *arXiv preprint arXiv:1708.06767*, vol. 4, no. 11, 2017.

[13] "Bimodal recurrent neural network for audiovisual voice activity detection multimodal signal processing (msp) laboratory," 2017.

[14] F. Tao and C. Busso, "Audiovisual speech activity detection with advanced long short-term memory," in *Interspeech*, 2018.

[15] T. Le Cornu and B. Milner, "Voicing classification of visual speech using convolutional neural networks," in *FAAVSP-The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*, 2015.

[16] R. Zazo, T. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform cldnns for voice activity detection," pp. 3668–3672, 09 2016.

[17] R. Aralikatti, D. K. Margam, T. Sharma, A. Thanda, and S. Venkatesan, "Global snr estimation of speech signals using entropy and uncertainty estimates from dropout networks," in *Proc. Interspeech 2018*, 2018, pp. 1878–1882. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1884

[18] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *Signal Processing Conference, 2008 16th European*. IEEE, 2008, pp. 1–5.

[19] F. Eyben, F. Weninger, S. Squartini, and B. W. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 483–487, 2013.

[20] S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," *Proc. AVSP2009*, pp. 151–154, 2009.

[21] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two novel visual voice activity detectors based on appearance models and retinal filtering," in *Signal Processing Conference, 2007 15th European*. IEEE, 2007, pp. 2409–2413.

[22] A. Aubrey, Y. Hicks, and J. Chambers, "Visual voice activity detection with optical flow," *IET image processing*, vol. 4, no. 6, pp. 463–472, 2010.

[23] C. T. Ishi, M. Sato, N. Hagita, and S. Lao, "Real-time audio-visual voice activity detection for speech recognition in noisy environments," in *Auditory-Visual Speech Processing 2010*, 2010.

[24] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[25] F. Tao, J. H. Hansen, and C. Busso, "An unsupervised visual-only voice activity detection approach using temporal orofacial features," *INTERSPEECH 2015*, vol. 2015, 2015.

[26] ——, "Improving boundary estimation in audiovisual speech activity detection using bayesian information criterion." in *INTERSPEECH*, 2016, pp. 2130–2134.

[27] B. Joosten, E. Postma, and E. Krahmer, "Visual voice activity detection at different speeds," in *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.

[28] R. Navarathna, D. Dean, S. Sridharan, C. Fookes, and P. Lucey, "Visual voice activity detection using frontal versus profile views," in *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*. IEEE, 2011, pp. 134–139.

[29] 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.

[30] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[31] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *2013 IEEE international conference on, Acoustics, speech and signal processing (ICASSP)*. IEEE, 2013, pp. 8614–8618.

[32] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[33] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," pp. 7378–7382, 10 2013.

[34] F. Patrona, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Visual voice activity detection in the wild," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 967–977, 2016.

[35] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *CoRR*, vol. abs/1612.08242, 2016. [Online]. Available: http://arxiv.org/abs/1612.08242

[36] C. Sanderson, "The vidtimit database," IDIAP, Tech. Rep., 2002.