INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Whisper activity detection using CNN-LSTM based attention pooling network trained for a speaker identification task

Abinay Reddy Naini[1], *Malla Satyapriya*[2], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
[2]Rajiv Gandhi University of Knowledge Technologies (RGUKT), Kadapa, 516330, India

nainireddy@iisc.ac.in, prasantg@iisc.ac.in

Abstract

In this work, we proposed a method to detect the whispered speech region in a noisy audio file called whisper activity detection (WAD). Due to the lack of pitch and noisy nature of whispered speech, it makes WAD a way more challenging task than standard voice activity detection (VAD). In this work, we proposed a Long-short term memory (LSTM) based whisper activity detection algorithm. However, this LSTM network is trained by keeping it as an attention pooling layer to a Convolutional neural network (CNN), which is trained for a speaker identification task. WAD experiments with 186 speakers, with eight noise types in seven different signal-to-noise ratio (SNR) conditions, show that the proposed method performs better than the best baseline scheme in most of the conditions. Particularly in the case of unknown noises and environmental conditions, the proposed WAD performs significantly better than the best baseline scheme. Another key advantage of the proposed WAD method is that it requires only a small part of the training data with annotation to fine-tune the post-processing parameters, unlike the existing baseline schemes requiring full training data annotated with the whispered speech regions.

Index Terms: Whisper activity detection, whispered speech, attention pooling network.

1. Introduction

Whisper activity detection (WAD) is the task of finding a whispered speech region in a noisy audio file. Whispered speech is one of the natural modes of speech production, primarily identified by lack of pitch [1, 2, 3]. Other factors that differentiate whispered speech from neutral speech include relatively flat spectral slope [4, 5], a considerable shift in lower-order formants [6]. A whispered speech sounds more like an unvoiced speech [3]. In typical voice activity detection (VAD), classifying unvoiced speech from noise is a more challenging task than detecting voiced speech [7, 8]. Detecting whispered speech accurately is useful for many applications. A robust WAD system can improve the performance of the speech recognition [9], speaker verification [10, 11] task based on whispered speech, by avoiding unnecessary processing of noise region. The recent advances in virtual assistant devices make them operate in whisper mode, which requires an accurate WAD to identify whispered speech regions in a conversation [12, 13].

Extensive research has been done on VAD. Among these, several new features have been proposed for VAD [14, 15]; on the other hand, most of the recent VAD works are based on Deep neural networks [16]. Despite the limited work, several attempts have been made in the literature to improve the WAD performance. Several features have been proposed to discriminate whispered speech from noise. These include spectral entropy [1], Mel-frequency cepstral coefficients [8], long-term spectral divergence [14], long-term signal variability (LTSV) [17, 18], auditory-inspired modulation spectrum [19], long-term logarithmic energy variation (LTLEV) [20], and fusion of group-delay-based subband modulation spectrum and correntropy features [21]. The LTLEV [20] based method showed better results compared to the other existing methods. However, existing WAD methods perform poorly in the presence of unseen noises and environmental conditions. The lack of the training data annotated with whispered speech regions, hinders development of WAD methods that require training with large amount of data.

To overcome these limitations, we, in this work, propose a Convolutional neural network- Long-short term memory (CNN-LSTM) based attention pooling network, which is trained for speaker identification. Recently, CNN based models showed the state of the art results for speaker verification task, particularly in low resource conditions [22]. This motivated us to consider the CNN layers for the primary network. In any neural network, a typical attention layer gives higher weightage to the region in the data, which contributes more to the network for target application [23]. It is well established in the literature that such attention layer in speech applications gives higher weightage to the speech regions than the noise regions of speech data [24]. We considered LSTM layers for the attention network to provide the full context for deciding the weights. Seyedmahdad et al. [24] showed that an attention layer helps in emotion recognition, by effectively weighing the speech region over a non-speech region of the data. Similarly, a typical attention based DNN model trained for neutral speaker identification can provide information for the speech region. However, it may not perform better compared to a model trained directly for VAD. But we find that attention weights from a CNN-LSTM based attention pooling model trained to classify speakers using whispered speech could do WAD better than the baseline method considered for WAD. The WAD experiments comprising 186 speakers' whispered speech data reveal that the proposed method performs better than the best baseline method.

2. Proposed WAD system

Fig 1 **(a)** shows the proposed CNN-LSTM attention model based WAD system. The CNN-LSTM based attention pooling network is trained for speaker identification task, as shown in Fig 1 **(b)**. Given an input speech, the output weights from the LSTM based attention block in Fig 1 **(b)** is passed through a post-processing block shown in Fig 1 **(c)** to make a WAD decision. The details of each of these blocks are given below.

2.1. Feature extraction

Given a speech signal, 24-dimensional log-filterbank features [22] are computed using a window of length T_w with a shift T_s.

http://dx.doi.org/10.21437/Interspeech.2020-3217

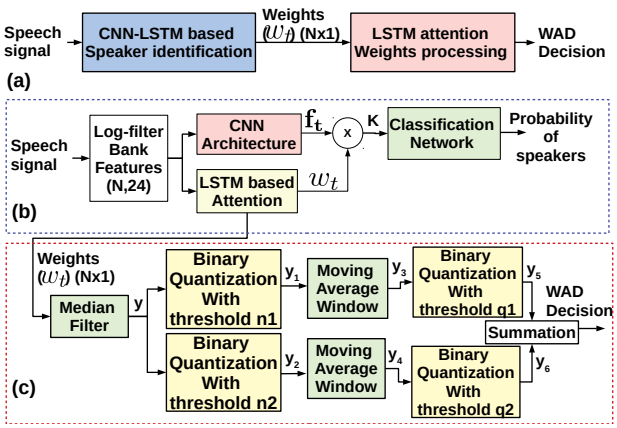

Figure 1: *Block diagram of the proposed CNN-LSTM attention pooling network based WAD system is shown in* ***(a)****, the block for CNN-LSTM based speaker identification is shown in* ***(b)****, the block for LSTM attention weights processing is shown in* ***(c)***.

Layer	parameters/Nodes
CNN Architecture	**(filter size, channels), pooling size**
1) CNN 1, Avg. pool 1 2) CNN 2, Avg. pool 2 3) CNN 3, Avg. pool 3	(3×3, 20), 2×1 (3×3, 20), 1×2 (3×3, 20), 2X1
LSTM based Attention	**Nodes, pooling size**
1) LSTM 1 2) LSTM 2, Avg. pool 4	24 1, 4×1
classification Network	
1) DNN 1 2) DNN 2	60 Num. Speakers

Figure 2: *Implementation details of the proposed CNN-LSTM attention pooling network*

2.2. Training of CNN-LSTM based speaker identification network

In the training phase, we considered the whispered speech data from the n speakers. In a CNN-LSTM attention pooling block, the LSTM based attention block computes the weights for N frames in the whispered speech file.

$$K = \sum_{t=1}^{N} w_t \mathbf{f}_t \quad (1)$$

Here the output from the LSTM block (w_t) weighs the output sequence from the CNN block ($\mathbf{f}_t$) to obtain weighted pooling vector K. The resulted pooled vector is given as an input to two fully connected DNN layers, known as the classification network. A categorical cross-entropy loss is computed at the end of the classification network, using the softmax activation for the n-class classification. This loss is backpropagated to train both the LSTM and the CNN blocks. The CNN block contains three sequential CNN layers with relu activation along with the average pooling layers. The LSTM based attention block contains two LSTM sequence-in and sequence-out layers along with an average pooling layer to match the number of frames of the CNN block output. The classification network block contains two fully connected layers with relu and softmax activation. The implementation details of the CNN-LSTM attention network are given in Fig 2.

2.3. Post-processing of LSTM attention weights for WAD decision

Given a test noisy whispered speech utterance, we compute features, which are passed as input to the LSTM block to obtain weights (between 0 and 1) for each frame. We have used these weights along with a post-processing block to make a WAD decision. There could be better post-processing methods that may require training with large annotated data. However, we have considered a simple approach requiring a small amount of annotated data. In this paper, we refer to the proposed Attention Pooling Weights based WAD as APW. Fig 1 **(c)** shows the post-processing block used for making a WAD decision. Details about all the sub-blocks are provided below.

2.3.1. Median filter

A median filter of M frames length is applied to the input weight vector (w_t) to smoothen the weight vector. This is done to avoid abrupt variations in the weight sequence. The median filter output is denoted by $\mathbf{y}$.

2.3.2. Binary Quantization (BQ)

The binary quantization block converts the median filtered weights to either 0 or 1 based on a given threshold. For example, the output, $\mathbf{y}_1$, of the BQ with threshold n1, for the input $\mathbf{y}$, is given by, $\mathbf{y}_1$(i) = 1, if $\mathbf{y}$(i)>n1, zero otherwise. Similarly $\mathbf{y}_2$, $\mathbf{y}_5$, $\mathbf{y}_6$ are also computed for the corresponding inputs, $\mathbf{y}$, $\mathbf{y}_3$, $\mathbf{y}_4$ with the thresholds n2, q1, and q2, respectively.

2.3.3. Moving Average Window (MAW)

In this block, a moving average window of size 1 second without any overlap is applied separately on $\mathbf{y}_1$ and $\mathbf{y}_2$, where each sample in the MAW is replaced with the average of all the samples in it. MAW block converts frame level decision to window level decision denoted by **y3** and **y4** corresponding to inputs **y1** and **y2**, respectively.

2.3.4. Summation

This is a frame-wise binary addition (logical OR) applied to $\mathbf{y}_5$ and $\mathbf{y}_6$. The output of this summation block is used for making a WAD decision, i.e., a frame with value one represents the whispered speech, and zero corresponds to the silence.

3. Experiments and results

3.1. Database

In this study, we have considered three different databases: (i) CHAINS [25] corpus contains 36 speakers, among them 28 are from the Eastern part of Ireland and the remaining eight speakers are from the UK and the USA. The CHAINS data was recorded in six different modes, including Synchronous, Fast, slow, whispered speech, etc. However, we have considered only whispered solo speech of 10 recordings from each of 36 subjects recorded at 44.1kHz. (ii) wTIMIT (whispered TIMIT) [26] dataset comprising 28 North American and 20 Singaporean

Database	Num. of Speakers		Recordings/speaker	
	Female	Male	Normal	Whisper
wSPIRE	41	61	-	50
wTIMIT	24	24	-	40
CHAINS	16	20	-	10

Table 1: *Number of male/female speakers and recordings per speaker for all three databases considered in this work.*

speakers, each speaking 450 utterances, both in whispered and neutral speech recorded at 44.1kHz. Although we have 450 utterances per speaker, we only considered 40 whispered speech utterances from each speaker to avoid class imbalance problem. (iii) wSPIRE database is an in-house recorded data containing 102 speakers, each speaking 50 sentences in neutral and whisper mode. This data is recorded parallelly across five recording devices. However, recordings from only two devices (ZOOM recorder [27], Moto G5 mobile) are considered for the experiments. These speakers are distributed across 13 dialect regions of India. We considered 50 whispered speech utterances from each of the speakers. Details of the number of male and female speakers and recordings per speaker are given in Table 1.

Although all three databases have both neutral and whisper, we considered only whispered data for this experiment. To experiment on different noise conditions, we considered the NOISEX-92 data set [28]. We considered a total of 8 different noises similar to the work by Nisha et al., [20], among which four were used for the training (namely Destroyer engine (DE), Factory (F), Machine (MC), and Pink noise (PN)). The other four (Hfchannel (HF), Machine gun (MG), Speech babble (SB), and white noise) were used for the testing. We resampled all the recordings to a sampling frequency (f_s) of 16kHz.

3.1.1. Baseline method (LTLEV)

The LTLEV based WAD [20] is used as the baseline for the experiments. We have not used the work by Raeesy et al. [13] as a baseline because our motivation is to detect whispered speech in high noise conditions, and, also, because of the non-availability of far-field whispered corpus. In the LTLEV method, the authors proposed a feature based on long-term log energy variation (LTLEV) [20]. In the training phase, given a noise condition, an optimal sub-band sequence is obtained using an automatic sub-band selection algorithm. In the test time, a given whispered utterance is filtered by the obtained optimal sub-band filter, then LTLEV features are computed using the filtered signal. Using a learned threshold with LTLEV, the WAD decision is obtained.

3.2. Experimental setup

We considered two experimental setups. In the first experimental setup (Ex-1), we have used only clean speech data for the training and tested on four noisy conditions with different levels of signal to noise ratio (SNR), namely, -10dB, -5dB, 0dB, 5dB, 10dB, 20dB, and 30dB. In the second experimental setup (Ex-2), we have used clean speech and their noisy versions for the training with additive training noises at SNRs , unseen to testing SNRs, namely, -5dB, 0dB, 5dB, 10dB, 15dB, 25dB to make the training robust to noise variations. Testing for Ex-2 is done in a maner similar to that for Ex-1.

For the training in the Ex-1 setup, we used 80 speakers from the wSPIRE database with 40 whispered speech utterances from each of them, as well as 40 whispered speech utterances from each of 48 speakers from the wTIMIT database. In the Ex-2 setup, we used the whispered utterances identical to those of Ex-1 but with noises added. The baseline method LTLEV requires optimal sub-band estimation for particular noise conditions. To compare the proposed method with the best possible combination in the baseline, we computed optimal sub-bands for all four training noises. For every test case, we computed performance using these four optimal sub-bands, and reported the best result for every test case. In the training stage, we consider only wSPIRE data recorded using a ZOOM recorder. Another parallel recording of the wSPIRE (Moto-g5) database is used for testing to evaluate the performance of the proposed (APW) and the baseline (LTLEV) methods on unseen device conditions. In the proposed method, 20% of the training data is used for the validation set.

For test data preparation, we combined ten utterances of whispered speech from a speaker with in-between silence of randomly chosen duration ranging from 1 to 2 seconds. Different test noises are added to this silence-appended speech file with varying levels of SNR. In the testing phase, we considered five different test cases indicated as **1A, 1B, 2, 3, 4** in both Ex-1 and Ex-2. These test cases indicate different testing conditions. In each test case, we have added four test noises (white, HF, MG, SB) with different levels of SNR. Table 2 shows the details of the testing conditions in each test case.

We extracted log-filterbank features using T_w = 25ms, and T_s = 10ms, and we have used $M = 10$ in the median filter. The thresholdings used in the post-processing step are obtained using the grid search to maximize accuracy on the validation set

Test cases	Testing conditons and (Data)
1A	Seen subject & environment (remaining 10 utterances from the 80 training speakers of wSPIRE)
1B	Unseen subject & seen environment (10 utterances from the remaining 22 speakers of wSPIRE)
2	Unseen subjects & environment (10 utterances of 36 speakers from the CHAINS database)
3	Seen subjects & unseen device (10 utterances of 80 training speakers from the wSPIRE recorded using Moto-g5)
4	Unseen subjects & device (10 utterances of 22 speakers from the wSPIRE recorded using Moto-g5)

Table 2: *Different testing conditions and the type of data used*

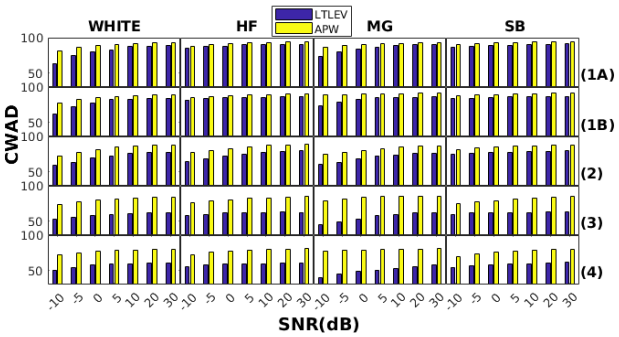

Figure 3: *Results for the experimental setup 1 (Ex-1), for all test cases including four testing noises with varying SNR values from -10dB to 30dB*

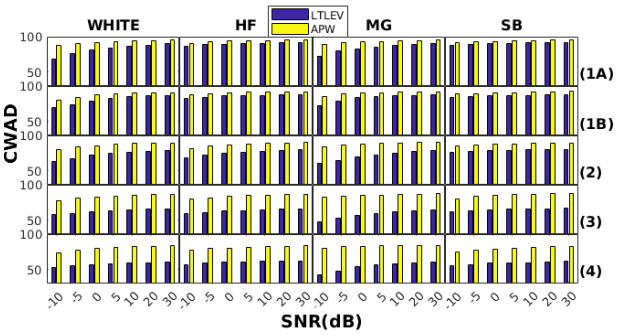

Figure 4: *Results for the experimental setup 2 (Ex-2), for all test cases including four testing noises with varying SNR values from -10dB to 30dB*

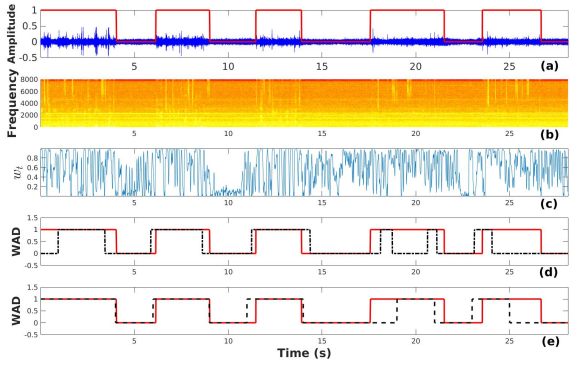

Figure 5: *A sample test file (a) waveform with added white noise at 0dB SNR to the first three test utterances, and at -10dB, SNR to the last two test utterances (b) corresponding spectrogram, (c) weights from the LSTM based attention block, (d) WAD decision, obtained from the Baseline method (black-dot-dashed), (e) WAD decision, obtained from the proposed method (black-dashed), for both (d,e) ground truth is in red.*

(20% of the training data), which are n1 = 0.91, n2 = 0.8, q1= 0.1, and q2 = 0.5. We have implemented the proposed network in tensorflow [29] and keras [30]. We have optimized the categorical cross entropy loss using adam optimizer [31], until the validation error increases.

In these experimets, we have used cumulative accuracy (ratio of the number of correctly classified frames to the total number of frames) to measure the performance, which is denoted as CWAD [20].

3.3. Results and discussion

Fig 3 shows the CWAD from the Ex-1 setup with all combinations of test cases with testing noises added at varying SNR values. We can observe from Fig 3 that baseline LTLEV performs better in test cases **1A, 1B**, compared to the remaining test cases, where the testing environment is unseen to training. However, in almost all cases, the proposed APW method showed an improvement over the baseline. The CWAD value from LTLEV for test cases **3, 4** is between $\sim$ 39% to 61.2%, which is a very low for two-class classification. This is probably due to the fact that the optimal sub-bands for the training and testing conditions are different. Hence, the baseline LTLEV does not perform well under unseen noises and environmental conditions. On the other hand, the performance drop from the **1A, B** to **3, 4** for the proposed APW is less compared to that for LTLEV making the proposed WAD method robust to different testing conditions. Particularly, for high noise conditions such as at -10dB SNR, APW based WAD showed an improvement of $\sim$9%, $\sim$10%, $\sim$12%, $\sim$19%, $\sim$21% over the LTLEV for the test types **1A, B, 2, 3, 4**, respectively, averaged across all testing noises.

Fig 4 shows the CWAD from the Ex-2 setup with all combinations of test cases with testing noises added at varying SNR values. Even after considering the four different sets of optimal sub-bands and retaining the best combination, the performance improvement showed by the baseline is not significant over that in Ex-1. However, the augmentation of noises in the training data resulted in a decent improvement for APW. Particularly, for **1A** test case, CWAD value reached as high as $\sim$96%. For the Ex-2 test cases **3, 4**, APW showed an improvement of $\sim$3.6%, $\sim$4.2% compared to Ex-1. At high SNR such as 30dB, for test cases **1A, B**, where the training testing conditions are similar, LTLEV performs on par with the proposed APW.

Among the different testing noise conditions, LTLEV performed better on SB, HF noises. On the other hand, the proposed APW performed the best for MG noise, and the lowest performance for SB noise. This could be due to the nonstationary nature of the MG noise, where the LSTM block learned to avoid these regions. With the noise augmentation during training, the proposed APW method showed the maximum average improvement of $\sim$7% for the white noise. The difference in the performance of the proposed and baseline methods is observed for MG noise testing at -10dB, which is as high as $\sim$22% for test type **4**.

Fig 5 shows a sample prediction by the proposed method as well as the baseline scheme. In this, the third row shows the weight sequence obtained from the LSTM network. We can observe from the Fig 5 that the proposed method's prediction is very close to the ground truth, for an SNR of 0dB (first three utterances in the illustrated example). The prediction deviates slightly from the ground truth for the low SNR case (last two utterances in the speech file, which is -10dB). However, the proposed method showed significant improvement over LTLEV even at a very low SNR, such as -10dB. This is evident by comparing Fig 5(d) and 5(e).

4. Conclusion

In this work, we proposed a whisper activity detection (WAD) method using the attention weights obtained from a CNN-LSTM attention pooling network trained for speaker identification. This is the first attempt on WAD using the neural networks based methods, and the proposed method outperformed the state-of-the-art results for the WAD in different additive noises and environmental conditions. The proposed method's key advantages are its robustness towards the varying noise and recording conditions and its ability to train on limited annotated data. In contrast, the existing baselines require the training data to be annotated with whispered speech regions. Thus the performance of the proposed method can be further improved with the addition of annotated data. For future work, we want to observe how the proposed method performs when trained for other speech tasks, such as gender classification.

Acknowledgement: We thank the Department of Science & Technology, Government of India for their support.

5. References

[1] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *Eighth Annual Conference of the International Speech Communication Association*, 2007, pp. 2289–2292.

[2] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7-8, pp. 515–520, 2002.

[3] V. C. Tartter, "What's in a whisper?" *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.

[4] X. Fan and J. H. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition," in *Ninth Annual Conference of the International Speech Communication Association*, 2008, pp. 1313–1316.

[5] ——, "Speaker identification within whispered speech audio streams," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1408–1421, 2011.

[6] S. T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels," *Acta Acustica united with Acustica*, vol. 84, no. 4, pp. 739–743, 1998.

[7] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Seventh European Conference on Speech Communication and Technology*, 2001, pp. 1887–1890.

[8] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *Proceedings of TENCon'93. IEEE Region 10 International Conference on Computers, Communications and Automation*, vol. 3, pp. 321–324.

[9] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[10] A. R. Naini, M. Achuth Rao, and P. K. Ghosh, "Whisper to neutral mapping using cosine similarity maximization in i-vector space for speaker verification," *Proc. Interspeech 2019*, pp. 4340–4344, 2019.

[11] A. R. Naini, A. Rao MV, and P. K. Ghosh, "Formant-gaps features for speaker verification using whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6231–6235.

[12] M. Cotescu, T. Drugman, G. Huybrechts, J. Lorenzo-Trueba, and A. Moinet, "Voice conversion for whispered speech synthesis," *IEEE Signal Processing Letters*, vol. 27, pp. 186–190, 2020.

[13] Z. Raeesy, K. Gillespie, C. Ma, T. Drugman, J. Gu, R. Maas, A. Rastrow, and B. Hoffmeister, "LSTM-based whisper detection," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 139–144.

[14] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[15] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[16] G. Gelly and J.-L. Gauvain, "Optimization of RNN-based speech activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 646–656, 2017.

[17] A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, and S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection." in *Interspeech*, 2013, pp. 718–722.

[18] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2010.

[19] M. Sarria-Paja and T. H. Falk, "Whispered speech detection in noise using auditory-inspired modulation spectrum features," *IEEE Signal Processing Letters*, vol. 20, no. 8, pp. 783–786, 2013.

[20] G. N. Meenakshi and P. K. Ghosh, "Robust whisper activity detection using long-term log energy variation of sub-band signal," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1859–1863, 2015.

[21] J. Wang, Y. Shang, S. Jiang, D. Gowda, and K. Lv, "Whispered speech detection using fusion of group-delay-based subband modulation spectrum and correntropy features," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1042–1046, 2016.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[23] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "ABCNN: Attention-based convolutional neural network for modeling sentence pairs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 259–272, 2016.

[24] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.

[25] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: Characterizing individual speakers," in *Proc of SPECOM*, vol. 6, 2006, pp. 431–435.

[26] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2011.

[27] "Zoom h6 handy recorder," 2018. [Online]. Available: https://www.zoom-na.com/products/field-video-recording/field-recording/zoom-h6-handy-recorder-0

[28] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[29] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[30] F. Chollet *et al.*, "Keras," 2015.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.