DUAL ATTENTION POOLING NETWORK FOR RECORDING DEVICE CLASSIFICATION USING NEUTRAL AND WHISPERED SPEECH

Abinay Reddy Naini[1], *Bhavuk Singhal*[2] *and Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science, Bangalore, India
[2]Information Technology, Bundelkhand Institute of Engineering and Technology, Jhansi, India

ABSTRACT

In this work, we proposed a method for recording device classification using the recorded speech signal. With the rapid increase in different mobile and professional recording devices, determining the source device has many applications in forensics and in further improving various speech-based applications. This paper proposes dual and single attention pooling-based convolutional neural networks (CNN) for recording device classification using neutral and whispered speech. Experiments using five recording devices with simultaneous direct recordings from 88 speakers speaking both in neutral and whisper and recordings from 21 mobile devices with simultaneous playback recordings reveal that the proposed dual attention pooling based CNN method performs better than the best baseline scheme. We show that we achieve a better performance in recording device classification with whispered speech recordings than corresponding neutral speech. We also demonstrate the importance of voiced/unvoiced speech and different frequency bands in classifying the recording devices.

***Index Terms*—** Recording device, whispered speech, Dual attention pooling network

1. INTRODUCTION

One of the crucial pieces of information in a recorded speech signal is the information about the device which was used to record the speech [1]. With a rapid increase in the number of different mobile and recording devices, identifying the recording device from a speech signal has various digital forensics and speech-enabled applications [2]. For example, establishing the source of audio or multimedia evidence in the court of law increases the authenticity of the evidence [2, 3]. This is because digital audio technology, at present, has facilitated the processing, manipulation and editing of audio by using sophisticated tools and software without leaving any perceptible trace [4]. Furthermore, knowing recording device characteristics has the potential to help other critical speech applications such as speech recognition and speaker verification by normalizing recording devices' variability.

Whispered speech is one of the natural modes of speech primarily characterized by lack of pitch [5,6]. Speakers often use whispered speech in their day-to-day life. With the recent advancements in speech-enabled and virtual assistant devices [7], robust methods for whispered speech detection [8], recognition, and whispered speech-based speaker verification [9, 10] had been developed. This motivated us to consider whispered speech together with neutral speech for recording device classification.

Research in recording device classification started with the success in camera identification from the produced image [11]. Several methods have been proposed in the literature to classify recording devices. We can broadly divide these methods into three categories. In the first category, researchers explored different features with the traditional statistical processing for the recording device classification. For example, Kraetzer et al. [12] used time, frequency, and cepstral domain features to classify four recording devices. Hanilci et al. [3] used Mel frequency cepstral coefficients (MFCC) with vector quantization (VQ) and support vector machines (SVM) classifiers. Hanilci et al. [2] also explored linear frequency cepstral coefficients with SVM. Pandey et al. [13] have used the estimate of the speech-free regions' power spectral density for recording device classification. Aggarwal et al. [14] proposed an MFCC feature extraction from an estimated noisy region of the speech. Jahanirad et al. [4, 15] investigated the use of entropy of Mel-cepstrum coefficients from near-silent segments. Anshan et al. [16] used device self-noise estimated from the silent segments. Further, Linear frequency cepstral coefficients [17, 18], Multitaper MFCC Features [19], random spectral features, and labeled spectral features [20] have also been explored.

In the second category, researchers explored different statistical machine learning approaches using MFCCs to classify recording devices. For example, Ling et al. [21] used Gaussian super vectors. Ling et al. [22] also used Kiss metric learning-based similarity matching on sparsely represented features for recording device verification. Jiang et al. [17] used Weighted Support Vector Machine with Weighted Majority Voting. In the third category, different deep features and deep learning methods have been explored. Yanxiong et al. [23] used features extracted using a deep auto-encoder network for recording device classification. Verma et al. [24, 25] proposed convolutional neural network (CNN)-based classification with absolute discrete Fourier transform (DFT) features. A detailed review of the recording device classification literature can be found in [4, 24]. Among these existing methods, CNN-based classification with absolute DFT features proposed by Vinay et al. [24] showed the best result. However, we observed a significant margin for improvement in recording device classification accuracy. Also, there is no work on recording device classification based on whispered speech to the best of our knowledge.

Considering the above limitations, we in this work proposed a single-attention pooling network (SA-CNN) and a dual-attention pooling network (DA-CNN) for recording device classification. Recently, CNN-based models showed state-of-the-art results for recording device classification [24, 25], which motivated us to keep the primary network based on CNNs. We observed that not all parts

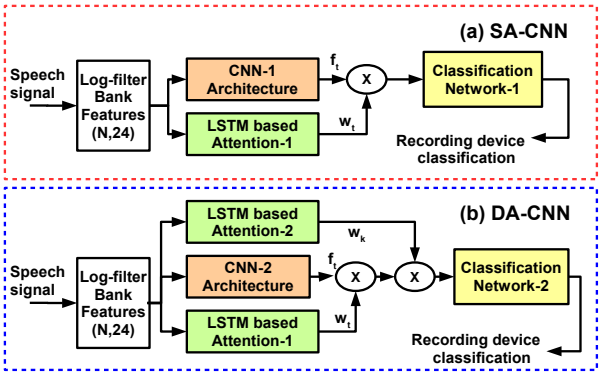


Fig. 1: Block diagram of the proposed CNN-LSTM attention pooling network for recording device classification. CNN-LSTM based Single-attention pooling network block is shown in **(a)**, the block for CNN-LSTM based dual-attention pooling network is shown in **(b)**.

of the speech spectrum contribute equally to the recording device classification. This motivated us to consider an attention-based CNN network in this work. Recently, CNN-based single and dual attention pooling networks showed promising results in many speech applications such as emotion recognition [26] and speaker verification [8]. Further, we observed that along with different sounds, different frequency bands in the spectrum contribute to different extent for recording device classification. This observation made us consider a dual attention pooling network apart from the single attention pooling network for both neutral and whispered speech. The recording device classification experiments comprising 21 recording devices with neutral speech and five recording devices with neutral and whispered speech revealed that the proposed dual and single attention pooling network-based methods perform better than the best baseline method.

2. PROPOSED METHOD

The block diagram shown in Fig. 1 explains both the proposed single attention pooling network (a) and the dual attention pooling network (b) for the proposed CNN-LSTM model. All the blocks are explained in detail below.

2.1. Feature extraction

24-dimensional log-filter bank features are obtained for a given speech signal using a window of length T_w with a shift of T_s.

2.2. Training of CNN-LSTM based single-attention pooling network (SA-CNN)

The CNN-LSTM based single-attention pooling network referred to as SA-CNN in this work is shown in Fig. 1(a). In the training phase, we consider neutral/whispered speech from n recording devices. Given the speech signal from any recording device, 24-dimensional filter bank features are computed and given as input to the combination of CNN-1 architecture and LSTM based attention-1. LSTM based attention-1 block computes weights (w_t), which weigh the output sequence consisting N frames from the CNN-1 architecture block (f_t) to obtain weighted embeddings ($\mathbf{K}$) as follows:

$$\mathbf{K} = \sum_{t=1}^{N} w_t f_t \tag{1}$$

The weighted embeddings ($\mathbf{K}$) are given as an input to the Classification network-1, which contains an average pooling layer followed by two fully connected DNN layers. At the end of the network, categorical cross-entropy loss is computed using the softmax activation for n-class classification. This loss is backpropagated to train both the LSTM based attention-1 and CNN-1 architecture blocks. The implementation details of the SA-CNN are given in Table 1.

2.3. Training of CNN-LSTM based dual-attention pooling network (DA-CNN)

The CNN-LSTM based dual-attention pooling network referred to as DA-CNN in this work is shown in Fig. 1(b). Similar to the SA-CNN, for a given speech signal from any recording device, 24-dimensional filter bank features are computed. The obtained input features are given as input to the LSTM based attention-1, LSTM based attention-2, and the CNN-2 architecture blocks. Similar to the SA-CNN, LSTM based attention-1 block computes weights (w_t^k), which weigh the output sequence from the CNN-2 architecture ($f_t^{'}$) to obtain weighted embeddings ($\mathbf{K}$) as shown in eq. 1. Further, a 24-dimensional weight vector (w_k) is obtained from LSTM based attention-2 block to further weigh the weighted embedding ($\mathbf{K}$) from each frame, as follows:

$$\mathbf{K}^{'} = \sum_{k=1}^{24} \left(\sum_{t=1}^{N} w_t^k f_t^k \right) w_k \tag{2}$$

The obtained embeddings ($\mathbf{K}^{'}$) are given as inputs to the Classification network-2. At the end of the network, categorical cross-entropy loss is computed using the softmax activation for n-class classification. This loss was backpropagated to train the LSTM based attention-1, LSTM based attention-2, Classification network-2, and CNN-2 architecture blocks. The implementation details of the DA-CNN are given in Table 1.

3. EXPERIMENTS AND RESULTS

3.1. Database

In this study, we have considered two databases:

MOBIPHONE dataset [27]: a publicly available dataset that contains recordings from 21 mobile phones of various models from seven different producers (HTC, LG, Sony, Nokia, Apple, Samsung, Vodafone). This dataset contains playback recordings of 12 male and 12 female speakers randomly chosen from the TIMIT database. All ten utterances from each speaker are recorded at 21 mobile phones in similar conditions. The recordings were initially done in adapted multi-rate (AMR) format with a 16 kHz sampling rate and were later converted into Waveform Audio File (WAV) format.

Blocks	Layers	Parameters/Nodes
		(filter size, channels), pooling size
CNN-1 Architecture	1) CNN 1	(3×3, 20), 2×1
	2) CNN 2	(3×3, 20), 1×2
	3) CNN 3	(3×3, 20), 2×1
CNN-2 Architecture	1) CNN 1	(3×3, 40), 2×1
	2) CNN 2	(3×3, 40), 1×1
	3) CNN 3	(3×3, 40), 2×1
		Nodes, pooling size
LSTM based Attention-1	1) LSTM 1	24, None
	2) LSTM 2	1, 4×1
LSTM based Attention-2	1) LSTM 1	24, None
	2) LSTM 2	24, 4×1
classification Network-1	1) DNN 1	60
	2) DNN 2	Num. devices
classification Network-2	1) DNN 1	120
	2) DNN 2	Num. devices

Table 1: Implementation details of the proposed CNN-LSTM based single and dual attention pooling network

wSPIRE dataset [28]: an in-house recorded data containing 88 speakers, each speaking 50 sentences in neutral and whisper mode. This data is recorded parallelly across the following five recording devices:

- Zoom H6 Handy Recorder with XYH-6 adjustable [29].
- Philips Stereo Headphones SHP-1900/97 [30].
- Apple iPhone 7 [31] (mobile phone).
- Nokia 5.1 [32] (mobile phone).
- motorola moto e5 plus [33] (mobile phone).

During the recording of the wSPIRE database, we placed all the mobile devices equidistantly from the speaker. Zoom H6 recorder, supported by a tripod, was placed in the right-front of the speaker. To indicate the beginning of the recording session, a 1-second long 1kHz tone was played, which also helped us synchronize the audio recordings in all the recording devices. All the recording devices used a sampling frequency of 44.1kHz, and later it is downsampled to 16kHz. All the speakers were either graduate/interning students or employees at the Indian Institute of Science, Bangalore. Details of the number of male and female speakers and recordings per speaker are given in Table 2.

3.2. Baseline method

We have used the method proposed by Vinay et al. [24] as the baseline in our experiments. In this method, the authors proposed a CNN architecture that was used to classify audio recordings from 19 different devices. A 4001-dimensional feature vector obtained using absolute values of the DFT of the recorded audio of 0.5 seconds duration is used as the input to the CNN architecture.

Database		wSPIRE	MOBIPHONE
Num. of Speakers	Female	34	12
	male	54	12
Recordings/ speaker	Neutral	50	10
	Whisper	50	-
Num. of rec. devices	Neutral	5	21
	Whisper	5	-

Table 2: Number of male/female speakers and recordings per speaker for both the databases considered in this work.

3.3. Experimental setup

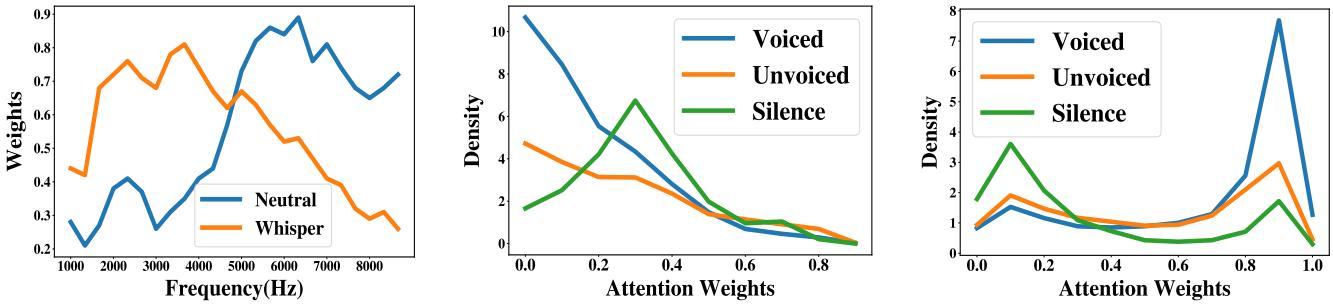
In the experimental setup, we have divided the data into train and test sets to perform three-fold cross-validation. We have performed a total of three experiments. In the first experiment (Ex-1), we have trained and tested the proposed and the baseline models on the MOBIPHONE dataset, which contains only neutral speech. In the second experiment, we have trained and tested the proposed and the baseline models on the wSPIRE dataset. We have further divided Ex-2 into four experiments based on the mode of data used for training and testing all the models. In the first part of Ex-2, we have used only the neutral speech of the wSPIRE dataset for all the models. In the second part of Ex-2, we have used only the whispered speech of the wSPIRE dataset for all the models. In the following two parts of Ex-2, we have trained the proposed and the baseline methods on one mode of speech and tested them on the other mode of speech (mismatched condition).

We have divided 88 speakers from the wSPIRE dataset into three sets having of 29, 29, 30 randomly chosen speakers. The MOBIPHONE dataset consisting of 24 speakers is split into three sets having eight speakers each, for three-fold cross-validation. No overlap between training and testing speakers is considered for the experiments. For mismatched train and test conditions in Ex-2, we tested on the same third fold using the mismatched mode of speech. For all test cases, we considered utterance level accuracy, that is, the number of correctly classified utterances to the total number of utterances to measure the performance. This is unlike the evaluation metric used in the baseline method, where the multiple utterance decisions are combined to make a single decision.

We extracted log-filterbank features using T_w = 25ms, and T_s = 10ms. We have implemented the proposed network in TensorFlow [34] and keras [35]. We have optimized the categorical cross entropy loss using adam optimizer [36], until the validation error increases.

3.4. Results and discussion

Table 3 shows a comparison of the accuracy in recording device classification for the proposed and the baseline methods in different experimental conditions. Among the proposed methods (SA-CNN, DA-CNN), the DA-CNN method achieved higher accuracy in all experimental conditions than the SA-CNN. Better performance of DA-CNN indicates that not all frequency bands are equally important in a recording device classification. In both Ex-1, Ex-2 the proposed SA-CNN showed an improvement of $\sim 4.2\%$, $\sim 7.4\%$ compared to that of the baseline method using wSPIRE neutral and whispered speech, respectively. The proposed DA-CNN showed a further improvement of $\sim 1.7\%$, $\sim 0.4\%$ over the SA-CNN. In Ex-



(a) (b) (c)

Fig. 2: **(a)** shows the 24 frequency bank weight distribution from DA-CNN for both neutral and whispered speech, **(b)** shows the distribution of weights from DA-CNN for the voiced, unvoiced, and silence regions of the neutral speech-based model, **(c)** shows the distribution of weights from DA-CNN for the voiced, unvoiced, and silence regions of the whispered speech-based model

Method	Ex-1	Ex-2			
	MOBI	wSPIRE			
	Neutral	Neutral	Whisper	N/W	W/N
Baseline	92.21	93.8	91.54	73.72	66.86
SA-CNN	86.64	97.82	99.08	60.52	49.36
DA-CNN	94.56	99.12	99.37	69.45	51.4

Table 3: Comparison of recording device classification accuracy for proposed and baseline methods in different experimental conditions for both neutral and whispered speech. MOBI represent MOBIPHONE dataset, N/W represent the convention that neutral speech is used for the training and whispered speech for the testing.

1, classification accuracies were higher for the wSPIRE dataset than the MOBIPHONE dataset because of the higher number of classes in the MOBIPHONE dataset. In Ex-1, the proposed DA-CNN showed an improvement over both the proposed SA-CNN and the baseline methods for the MOBIPHONE dataset. It could be due to an increase in the importance of having dual attention with the number of recording devices.

It is interesting to see that recording device classification using whispered speech resulted in a better performance than neutral speech for both the proposed methods, unlike the baseline method. This indicates that the recording device information is not equally masked in the whispered speech's voiced, unvoiced, silent regions. Further, using the attention network helped the proposed methods to focus on un-masked regions to obtain better classification accuracy.

From Table 3, we can also see that recording device classification accuracy in Ex-2 (mismatched condition) was significantly lower than that with the matched train and test conditions (Ex-1, Ex-2). We can also observe that the baseline method has better accuracy than the proposed methods in Ex-2 mismatched conditions. This shows that the proposed methods were more adapted to the neutral/whisper mode of training speech and, hence, less generalizable across modes of speech. Further, it justifies the need for an attention layer in the proposed method for matched train and test conditions because different phonetic and frequency bands have different recording device information levels. An illustrative understanding of the importance of having dual attention is provided in Fig.2.

Fig.2(a) shows the 24 frequency bands weight (w_k) distribution obtained from the LSTM based attention-2 block of DA-CNN for both neutral and whispered speech of wSPIRE dataset in Ex-2. It is clear from the figure that the model trained using neutral speech weighs the upper half of the speech spectrum (primarily un-voiced region). On the other hand, the model trained using the whispered speech weighs the lower part (voiced region) of the speech spectrum. A similar observation is made from Fig.2(b), (c), which show the distribution of the weights (w_t) from the LSTM based attention-2 block of DA-CNN for the voiced, unvoiced, and silence regions of the neutral and whispered speech models correspondingly (Continuous value weights are quantized to 11 values, between zero and one, with steps of 0.1). We can observe from Fig.2(b) (corresponding to neutral speech) that silence and unvoiced regions have slightly higher weights than voiced regions. Similarly, from Fig.2(c) (corresponding to whispered speech), the voiced region has significantly higher weights than the unvoiced and silent regions.

4. CONCLUSION

In this work, we proposed a single and a dual-attention pooling based CNN-LSTM network for recording device classification. Experiments with both the wSPIRE and MOBIPHONE datasets revealed the importance of the attention layers in the proposed method. We also obtained the state-of-the-art results in recording device classification using the proposed dual-attention pooling based CNN-LSTM network. Further, this is the first attempt to classify recording devices using whispered speech data, which revealed that whispered speech is more suitable for recording device classification than the corresponding neutral speech. In future, we want to improve other speech applications such as speech recognition and speaker verification by including the recording device embeddings along with a regular feature vector. We also want to make the recording device classification robust to different noise conditions in future.

5. REFERENCES

[1] C. Hanilci, F. Ertas, T. Ertas, and Eskidere, "Recognition of brand and models of cell-phones from recorded speech signals," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 625–634, 2012.

[2] C. Hanilçi and T. Kinnunen, "Source cell-phone recognition from recorded speech using non-speech segments," *Digital Signal Processing*, vol. 35, pp. 75–85, 2014.

[3] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertaş, "Speaker identification from shouted speech: Analysis and compensation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8027–8031.

[4] M. Jahanirad, A. W. A. Wahab, N. B. Anuar, M. Y. I. Idris, and M. N. Ayub, "Blind source mobile device identification based on recorded call," *Engineering Applications of Artificial Intelligence*, vol. 36, pp. 320–331, 2014.

[5] 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.

[6] C. Zhang, G. S. Morrison, and P. Rose, "Forensic speaker recognition in chinese: a multivariate likelihood ratio discrimination on /i/and/y," in *Ninth Annual Conference of the International Speech Communication Association*, 2008, pp. 1937–1940.

[7] 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.

[8] A. R. Naini, M. Satyapriya, and P. K. Ghosh, "Whisper activity detection using CNN-LSTM based attention pooling network trained for a speaker identification task," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020. International Speech Communication Association, 2020, pp. 2922–2926.

[9] 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.

[10] 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.

[11] M. Darvish Morshedi Hosseini and M. Goljan, "Camera identification from hdr images," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2019, pp. 69–76.

[12] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: a first practical evaluation on microphone and environment classification," in *Proceedings of the 9th workshop on Multimedia & security*, 2007, pp. 63–74.

[13] V. Pandey, V. K. Verma, and N. Khanna, "Cell-phone identification from audio recordings using psd of speech-free regions," in *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science*. IEEE, 2014, pp. 1–6.

[14] R. Aggarwal, S. Singh, A. K. Roul, and N. Khanna, "Cellphone identification using noise estimates from recorded audio," in *2014 International Conference on Communication and Signal Processing*. IEEE, 2014, pp. 1218–1222.

[15] M. Jahanirad, N. B. Anuar, and A. W. A. Wahab, "Blind source computer device identification from recorded voip calls for forensic investigation," *Forensic science international*, vol. 272, pp. 111–126, 2017.

[16] A. PEI, R. WANG, and D. YAN, "Cell-phone origin identification based on spectral features of device self-noise," *Telecommunications Science*, vol. 33, no. 1, pp. 85–94, 2017.

[17] Y. Jiang and F. H. Leung, "Mobile phone identification from speech recordings using weighted support vector machine," in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2016, pp. 963–968.

[18] D. Garcia-Romero and C. Y. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 1806–1809.

[19] Ö. Eskidere and A. Karatutlu, "Source microphone identification using multitaper MFCC features," in *2015 9th International Conference on Electrical and Electronics Engineering (ELECO)*. IEEE, 2015, pp. 227–231.

[20] Y. Panagakis and C. Kotropoulos, "Telephone handset identification by feature selection and sparse representations," in *2012 IEEE International workshop on information forensics and security (WIFS)*. IEEE, 2012, pp. 73–78.

[21] L. Zou, Q. He, and X. Feng, "Cell phone verification from speech recordings using sparse representation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1787–1791.

[22] L. Zou, Q. He, J. Yang, and Y. Li, "Source cell phone matching from speech recordings by sparse representation and kiss metric," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2079–2083.

[23] Y. Li, X. Zhang, X. Li, Y. Zhang, J. Yang, and Q. He, "Mobile phone clustering from speech recordings using deep representation and spectral clustering," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 4, pp. 965–977, 2017.

[24] V. Verma and N. Khanna, "Speaker-independent source cell-phone identification for re-compressed and noisy audio recordings," *Multimedia Tools and Applications*, pp. 1–23, 2021.

[25] ——, "Cnn-based system for speaker independent cell-phone identification from recorded audio." in *CVPR Workshops*, 2019, pp. 53–61.

[26] 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.

[27] C. Kotropoulos and S. Samaras, "Mobile phone identification using recorded speech signals," in *2014 19th International Conference on Digital Signal Processing*. IEEE, 2014, pp. 586–591.

[28] B. Singhal, A. R. Naini, and P. K. Ghosh, "wspire: A parallel multi-device corpus in neutral and whispered speech," in *24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2021, pp. 146–151.

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

[30] "Philips stereo headphones," 2018. [Online]. Available: https://www.philips.co.in/c-p/SHP190097/$stereo-headphones$

[31] "Apple-iphone-7." [Online]. Available: https://support.apple.com/kb/SP743?locale=enIN

[32] "Nokia-5.1." [Online]. Available: https://www.nokia.com/phones/en$in/nokia-5-1$

[33] "motorola-moto-e5-plus." [Online]. Available: https://www.motorola.in/smartphones-moto-e5-plus/p

[34] 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.

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

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