INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Automatic Detection of Shouted Speech Segments in Indian News Debates

Shikha Baghel[1], *Mrinmoy Bhattacharjee*[1], *S. R. M. Prasanna*[2], *Prithwijit Guha*[1]

[1]Dept. of EEE, Indian Institute of Technology Guwahati, Guwahati-781039, India
[2]Dept. of EE, Indian Institute of Technology Dharwad, Dharwad-580011, India

shikha.baghel@iitg.ac.in, mrinmoy.bhattacharjee@iitg.ac.in, prasanna@iitdh.ac.in, pguha@iitg.ac.in

Abstract

Shouted speech detection is an essential pre-processing step in conventional speech processing systems such as speech and speaker recognition, speaker diarization, and others. Excitation source plays an important role in shouted speech production. This work explores feature computed from the Integrated Linear Prediction Residual (ILPR) signal for shouted speech detection in Indian news debates. The log spectrogram of ILPR signal provides time-frequency characteristics of excitation source signal. The proposed shouted speech detection system is deep network with CNN-based autoencoder and attention-based classifier sub-modules. The Autoencoder sub-network aids the classifier in learning discriminative deep embeddings for better classification. The proposed classifier is equipped with attention mechanism and Bidirectional Gated Recurrent Units. Classification results show that the proposed system with excitation feature performs better than baseline log spectrogram computed from the pre-emphasized speech signal. A score-level fusion of the classifiers trained on the source feature and the baseline feature provides the best performance. The performance of the proposed shouted speech detection is also evaluated at various speech segment durations.

Index Terms: Shouted speech detection, Indian news debate, CNN, attention, Bidirectional GRU, autoencoder, classifier.

1. Introduction

News debates have gradually become one of the main opinion-forming discussion-based news programmes in India. The motive of a news debate programme is to analyze, discuss and investigate the topics related to social, economical, political and cultural issues [1]. Each news debate consists of a panel of experts on the topic at hand. The panel discussions are aimed at making a viewer aware about various aspects of a trending news event in the world. Currently, broadcasting debates on popular opinion pieces that garner a huge emotional response from both the panelists and viewers alike is a common trend followed across most major news channels of India [1]. In such debates, shouting is frequently encountered. This work aims to automatically detect the presence of shouted speech in Indian news debates.

For any high-level application of news debates such as automatic content analysis, the first pre-processing step is to identify the complex segments from the audio signal and process them separately. One such complex speech event of an Indian news debate is shouted speech. Due to the frequent presence of shouted speech, Indian news debates are sometimes called as heated debate. Most of the conventional speech processing systems such as speaker diarization, speech and speaker recognition systems are trained on normal speech. However, the performance of such systems degrades when test data deviates from normally phonated speech [2] such as in shouted speech. Moreover, automatic shouted speech detection is also an important sub-task in many applications such as behavior studies, health monitoring, security, surveillance [3].

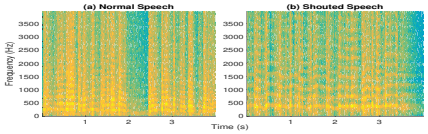

Figure 1: *Illustrating the ILPR log spectrogram for (a) normal and (b) shouted speech. More prominent harmonic patterns are observed in the ILPR log spectrogram for shouted speech in comparison to normal speech.*

1.1. Related Works

Shouted speech is defined as the speech with high vocal effort [4]. A high vocal-effort is associated with large activity of respiratory muscles, which involves pumping out a large volume of air from lungs [5]. The fundamental frequency (F_0, henceforth) has been extensively used for characterizing shouted speech [5, 6]. The changes in excitation source characteristics due to shouted speech have been studied by exploring different aspects of Electroglottogram (EGG, henceforth) and Differenced Electroglottogram (DEGG, henceforth) signals [5, 4, 7, 6]. The proportion of open and closed phase durations in a glottal cycle have been explored for shouted and normal speech in terms of closed quotient (CQ), open quotient (OQ), closing quotient (CIQ), and speed quotient (SQ) parameters [4, 7]. Baghel et al. [8, 9] proposed three DEGG based features viz. glottal open phase tilt, flatness of glottal cycle and open phase triangle area for capturing the changes in glottal cycle shape due to shouted speech. The effective decay time of a glottal cycle has been parameterized by normalized amplitude quotient for shouted and normal speech [6]. Baghel et al. [10] explored DCT-ILPR feature for capturing glottal cycle shape for shouted and normal speech classification task. They also analyzed spectral periodicity and smoothed spectral characteristics of LP residual signal in shouted speech detection.

Shouted speech production is generally associated with amplified articulatory movements such as lower jaw position, wide lip opening, tongue movement. This modifies the vocal-tract shape and hence results in a deviated vocal-tract characteristics for shouted speech. The Mel-Frequency Cepstral Coefficients (MFCC, henceforth) is the most extensively used feature for shouted speech characterization [11, 3, 12]. Spectral features such as spectral tilt, centroid, balance, flatness, and spread have also been used for shouted speech detection and related tasks [13]. Mittal and Vuppala [14] studied dominant frequency (F_D) for shouted speech and found a higher value of F_D for shouted speech in contrast to normal speech. The spectral har-

http://dx.doi.org/10.21437/Interspeech.2021-1592

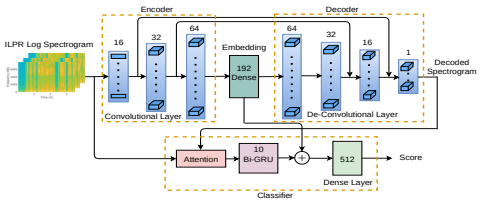

Figure 2: *Proposed shouted speech detection system.*

monic richness of shouted and loud speech is higher than that of soft and normal speech [4]. Few works have also studied Linear Prediction Coefficients (LPC, henceforth) and perceptual linear prediction coefficients (PLP) for shouted speech detection [15].

Gaussian Mixture Model (GMM) [16] and Support-Vector Machine (SVM) [9, 17] based classifiers have been used in some works for shout detection task. In recent years, deep learning based approaches such as Deep Neural Network (DNN, henceforth) [10], 1D-Convolutional Neural Network (CNN, henceforth) [18] and Deep Belief Networks (DBN)-DNN [19] have been used for shout detection task.

1.2. Motivation and Contribution

Shouted speech detection has mostly been attempted using vocal tract representations such as MFCC. Excitation source based works have mostly explored DEGG based features for shouted speech characterization. However, DEGG may not be available in many practical applications. Hence, the equivalent representation needs to be extracted directly from speech signal. In this direction, Integrated Linear Prediction Residual (ILPR, henceforth) [20] is used in this work. The time domain glottal cycle shape information has been explored in [10] for shouted speech (recorded under controlled environment) detection. However, the present work explores the time-frequency characteristics of ILPR signal.

Contributions of the present work are three folds. First, this work is an attempt to identify one of the complex and frequently occurring speech event, viz. shouted speech in Indian news debate audio. Most of the existing works used speech signals recorded either in a controlled (mostly) or in an open (few) environment where speakers are instructed to produce shouted speech. Hence, synthetic stimuli were used to produce shouted speech for such works. However, news debate audio contains naturally produced shouted speech. Moreover, there are many other complexities present in news debate data, such as varying recording conditions. Second, the present work explores the time-frequency characteristics of excitation source signal for shouted speech detection system using ILPR representation. Third, a novel classifier for shouted speech detection system is proposed in this study. The proposed system comprises of two sub-modules, an autoencoder and a classifier. The motive of incorporating autoencoder is to learn generative representation which can later help in data augmentation for shouted speech. The architecture of autoencoder is fully-convolutional. The classifier sub-module contains attention mechanism, Bidirectional Gated Recurrent Units (Bi-GRU, henceforth), and a fully connected (FC, henceforth) layer. The attention module provides a weighted spectrogram based on the similarity with decoded spectrogram. This can help in training a better classifier. The Bi-GRU is used to learn the temporal sequence of the weighted spectrogram. The embedding of spatial information generated by the autoencoder from time-frequency representation is used in conjunction with the temporal embedding obtained from Bi-GRU to perform the classification.

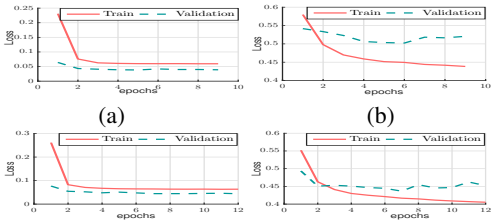

Figure 3: *Depicting training and validation losses of the proposed system. ILPR-Log spectrogram: (a) autoencoder loss, (b) classifier loss. Log-spectrogram: (c) autoencoder loss, (d) classifier loss.*

The paper is organised as follows. The proposed approach is described in section 2. The news debate dataset is briefly explained in section 3. Section 4 illustrates and discusses the experimental results obtained in the present work. Finally the work is concluded in section 5.

2. Proposed Approach

This section briefly describes the proposed approach for detecting shouted speech segments in news debates.

2.1. Integrated Linear Prediction Residual (ILPR)

A faster vibration and abrupt closing of vocal folds lead to a deviated glottal cycle shape for shouted speech in comparison to normal speech. These deviations are expected to be reflected in frequency domain in terms of prominent harmonics in the ILPR spectrogram. Figure 1 illustrates the ILPR Log spectrogram for normal and shouted speech of the same speaker (male). The prominent harmonics can be observed for shouted speech (Figure 1 (b)) than normal speech (Figure 1 (a)). The speech signal $s[n](n = 0, \ldots N)$ is pre-emphasized to suppress the effect of excitation source. The pre-emphasized speech $s_p[n]$ is used to obtain LPCs values using LP analysis. Now, the original speech signal $s[n]$ (instead of $s_p[n]$) is passed through an inverse filter (using LPCs) to obtain ILPR signal [20]. Further, each frame of ILPR signal is transformed to frequency domain using Discrete Fourier Transform (DFT). The logarithm of magnitude spectra is considered as the ILPR Log spectrogram.

2.2. Shouted Speech Detection System

The proposed shouted speech detection system consists of two sub-modules, viz. autoencoder and classifier. The goal behind this architecture is to build a model which has both generative and discriminative capabilities. The proportion of normal and shouted speech is not equal in an Indian news debate audio. This data imbalance between the classes may impose a limitation on further processing of debate data for high-level applications such as content analysis. Hence, a generative model like an autoencoder can be useful in data augmentation for shouted speech. The proposed classifier architecture is illustrated in Fig. 2. The proposed design is discussed in detail in the following text.

The proposed autoencoder architecture contains convolutional (encoder block) and deconvolutional layers (decoder block) with skip connections. The skip connections between encoder and decoder are used to ensure the proper gradient propagation throughout the network. There is an embedding layer (FC layer with 192 nodes) between the encoder and decoder block. The embedding layer of the autoencoder encapsulates

Table 1: *Classification results. LS stands for Log-Spec and ILS stands for ILPR-Log Spec. Results are reported as* $\mu \pm \sigma$.

Measures		LS	ILS	ILS + LS
Overall accuracy		75.52 $\pm$ 1.85	77.01$\pm$1.42	**78.01** $\pm$ **1.56**
F1-score	Normal	80.60 $\pm$ 3.91	82.22 $\pm$ 1.98	**82.81** $\pm$ **2.59**
	Shouted	70.69 $\pm$ 4.56	69.41 $\pm$ 4.78	**71.55** $\pm$ **4.67**
	Average	75.64 $\pm$ 3.76	75.82 $\pm$ 2.93	**77.18** $\pm$ **3.41**

the spatio-temporal information of the spectrogram. The input to the encoder is the spectrogram of 500ms speech signal. The encoder block consists of three convolutional layers with 16, 32 and 64 number of kernels, respectively. The output of each convolutional layer is batch normalized. Maxpooling with a pool size of $(2, 2)$ and stride of $(2, 2)$ is performed. The flattened output of last convolutional layer of encoder is fed to the embedding layer with 192 nodes. The output of embedding layer is passed to the decoder block. The decoder comprises three deconvolutional layers with 64, 32 and 16 number of kernels, respectively. A kernel size of $(3, 3)$ with a stride of $(2, 2)$ is used in convolutional and deconvolutional layers of encoder-decoder. To map the output of decoder in the same size as that of the input spectrogram, a convolutional layer with one kernel is used at the end of the decoder. All the layers in the autoencoder sub-module are non-linearly activated with Rectified Linear Units. The layer weights are $l2$ regularized and dropouts are applied after each layer to avoid overfitting of the network.

The classifier sub-module consists of an attention block, a Bi-GRU and an FC layer. The attention block takes two inputs, viz. original spectrogram and decoded autoencoder output, and performs weighing of the spectrogram frames based on the current context. The attention block calculates the weights according to the similarity between the original and decoded spectrograms. The calculated weights are multiplied with the original spectrogram to obtain a weighted spectrogram at the output of the attention block. The weighted spectrogram is passed through a Bi-GRU (10 nodes in each direction) to learn its temporal sequence. The embedding obtained from Bi-GRU containing temporal context information is concatenated with the embedding obtained from the autoencoder. The motive behind this concatenation is to utilize spatial and temporal context information for the classification task. The concatenated embedding is fed to an FC layer with 512 nodes. Finally, the scores for an input spectrogram is obtained from the output layer of the classifier that consists of a single sigmoid activated neuron. The proposed classifier is trained with a Nadam optimizer which is a variant of the popular Adam optimizer that uses Nesterov momentum. The initial learning rate is set as 0.0001. The autoencoder sub-module is trained using mean-squared error loss, while the classifier sub-module is trained using binary cross-entropy loss. A weighted sum of both the losses is used for the joint training of both the sub-modules. Since the goal of the current work is improved detection of shouted speech, more weight is given to the classifier sub-module's loss than that of the auto-encoder sub-module.

3. Indian News Debate Dataset

This work is evaluated on the speech signals extracted from 15 Indian English news debates. The audio signal of a news debate is broadly categorized into three categories, viz. normal speech, shouted speech and miscellaneous. The miscellaneous category includes music, speech with music, outside reporting, laugh, cough, and others. Audio signals are resampled at 16 kHz. The duration of a news debate typically varies from 20 min to 60 min. Processing such a long duration audio signal requires a lot of memory. Hence, each debate audio has been split into 10 min non-overlapping audio chunks. The annotation for a shouted speech is a subjective task. Soft-spoken persons may consider loud speech as shouted speech. However, the same loud speech segment might be a normal speech for another person. To avoid such subjective ambiguity, each 10 min audio chunk is annotated by three annotators, and the final label is decided using a majority vote. If all the three annotators have different labels (normal / shout/ miscellaneous) for a speech segment, then final label would be miscellaneous. The whole annotation process involves a total of 40 annotators. These annotators are the students of IITG studying at bachelor, masters and research level. There are many instances where a speaker gives high emphasis on certain words while speaking in normal vocal mode. Examples of such words and phrases are 'but', 'so what', 'and', 'especially', and so on. The speech signal characteristics for such words and phrases are expected to be inclined towards shouted speech. However, the usage of such words and phrases is in the context of normal speaking mode. Hence, these are marked as normal speech. Another complex situation is that when a speaker is speaking in shouted vocal mode, the volume of his/her microphone is lowered by the news channel administrators. Such segments are marked as shouted speech. There are cases where simultaneously two speakers are speaking. The dominant speaker is speaking in normal vocal mode, whereas the other speaker is shouting. The annotation of such segments is done based on the perception of the resultant speech. The news debates considered for this study are in Indian English language. However, there are instances where panel members speak in their native language, such as Hindi. Thus, some cases of code-switching are also present in news debate audio. Hence, the news debate dataset contains lots of complex scenarios, as mentioned in this section. The total duration of the data is $\approx$ 10 hours, which includes normal speech (4.8 hrs), shouted speech (3.14 hrs), and the remaining data belongs to the miscellaneous class. The normal speech and shouted speech data are used to evaluate the present work.

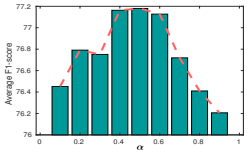

Figure 4: *Illustrating classification performance for feature combination at different* α *values.*

4. Experiments and Results

This section presents and discusses the experimental results of the present work. The analysis and detection of shouted speech have mostly been attempted by utilizing vocal tract representation of speech signal [11, 3, 12, 13, 14, 15]. Hence, this work uses the spectrogram, extracted from pre-emphasized speech signal, as the baseline method. The logarithm operation is performed on the magnitude spectrogram to obtain a log-spectrogram.

4.1. Classification using the Proposed Model

Evaluation of the proposed approach is performed on the speech data of 15 Indian English news debates. The whole dataset is split into three non-overlapping folds. Effectively, each fold contains speech signals corresponding to five debates. For one iteration, data of any two folds are considered for training, while

Table 2: *Classification results for* 1 *sec and* 1.5 *sec segment durations. LS stands for Log-Spec and ILS stands for ILPR-Log Spec.*

	1 sec				1.5 sec			
		F1-score ($\mu \pm \sigma$)				F1-score ($\mu \pm \sigma$)		
Features	**Accuracy ($\mu \pm \sigma$)**	**Normal**	**Shouted**	**Average**	**Accuracy ($\mu \pm \sigma$)**	**Normal**	**Shouted**	**Average**
LS	82.90 ± 1.46	85.72±2.33	78.46±0.39	82.09± 1.25	82.39 ± 1.41	85.27 ± 2.27	77.92±0.59	81.59 ± 1.23
ILS	83.24 ± 1.79	86.71±2.12	77.26± 2.07	81.98±2.04	82.76 ± 2.14	86.35± 2.44	76.54± 2.68	81.44±2.50
ILS + LS	**83.95 ± 1.03**	**87.01 ± 1.47**	**78.87 ± 1.24**	**82.94 ± 1.07**	**83.32 ± 0.99**	**86.50 ± 1.33**	**78.06 ± 1.62**	**82.28 ± 1.17**

the test set contains the remaining fold's data. The train set is further split into $70 : 30$ ratio and considered as training and validation sets, respectively. Therefore, three different combinations of training, validation, and test sets are obtained. The proposed approach is evaluated independently on each of these combinations. The performance is reported in terms of mean (μ) and standard deviation (σ) of performance of the three folds.

The perceptual understanding of shouted speech is better when a speech signal with sufficient duration is given. It is difficult for humans to detect shouted speech with a frame-length speech signal. Therefore, the present work reports the classification results over a segment duration of 500 ms. The speech signal is processed with a frame size of 25 ms with a shift of 10 ms. The input to the proposed system is a spectrogram of size $(200, 48)$, where 200 is the spectral resolution and 48 is the temporal resolution. Figure 3 illustrates the different losses encountered during the training of the model. Figure 3 (a) and (c) show the mean-squared loss of autoencoder for ILPR-Log spectrogram and Log-spectrogram, respectively. The training and validation losses stabilize after the 4^{th} epoch. The loss evolution plots indicate that the classifier sub-module tends to overfit slightly for both the features. The lack of enough data to learn all the complexities present in the news debate audio, as discussed previously, might be a reason for this. The autoencoder sub-module trains well, probably aided by proper gradient flow through the skip connections.

The classification results are reported in Table 1. The overall accuracy of ILPR-Log spectrogram is higher than the Log-spectrogram (baseline) by nearly 1.5%. The standard deviation (σ) is also lower than the baseline feature. The F1-score for normal class is also better with the proposed approach. However, the baseline has 1% higher F1-score for shouted speech. The baseline method has higher σ values for both the normal and shouted speech. In contrast, the proposed approach has comparatively lower σ value for the normal class. This signifies that the patterns learned by ILPR-Log spectrograms are more consistent across the dataset for normal speech. However, a higher σ value is observed for shouted speech for both the proposed and baseline approaches. The reason may be the high variability present in shouted speech in the news debate dataset. The F1-score of shouted speech is lower than that of normal speech. One possible reason for this could be the data imbalance between the classes. The overall F1-score for the proposed approach is slightly higher than the baseline. The classification performance of the proposed approach justifies the motivation of the present work. Hence, it can be said that the spectro-temporal patterns of ILPR signal contain enough class-specific information which can be utilized for shouted and normal speech classification task.

4.2. Excitation Source vs Vocal Tract Features

The proposed approach utilizes the excitation source characteristics for shouted speech detection. The baseline captures vocal tract (VT, henceforth) information of the speech signal. Both of the representations carry complementary information of the speech signal. Therefore, this work further explores the usage of ILPR-Log spectrogram in combination with Log-spectrogram to utilize both excitation and VT information for the current task. The features are combined at score level by considering equal weights for both the representations. The classification performance for the feature combination is reported in Table 1. The overall accuracy improved with the combination of ILPR-Log spectrogram and Log-spectrogram in comparison to individual features. Similarly, F1-scores of individual classes are also enhanced for the feature combination. Hence, it can be said that the usage of complementary information of excitation source and VT representation can yield a more discriminative shouted speech detection system. The score level fusion of features is further extended by analysing the effect of excitation source and VT feature by giving different weights (α). The features are combined as αLS + $(1 - \alpha)$ILS. Here, LS stands for Log-spectrogram and ILS represents ILPR-Log spectrogram. Figure 4 illustrates the classification performances for different α values varies from 0.1 to 0.9. The highest results are obtained for $\alpha = 0.5$. This shows that both VT and excitation source characteristics are equally important for the present task.

4.3. Effect of Segment Duration

The effect of segment duration on shouted speech detection is studied in this work. The classification is performed for two different segment durations, viz. 1 sec and 1.5 sec. The classification performance of different segment durations are reported in Table 2. A significant improvement is observed for both the features for 1 sec duration in comparison to 500 ms. The increment in F1-score is higher for shouted speech than that of normal speech. This trend is consistent with both the features. This signifies that the segment duration plays an important role in detecting shouted speech. The highest performance is obtained for the combination of ILPR-Log spectrogram and Log-spectrogram. The performance for 1.5 sec duration is comparable with that of 1 sec duration for individual features and their combination. Hence, it can be said that (nearly) 1 sec duration can be an appropriate choice of segment duration for detecting shouted speech.

5. Conclusions

This work proposes a novel shouted speech detection system. The time-frequency information of the ILPR signal is utilized for the current task. The proposed classifier has both generative (autoencoder) and discriminative (classifier) capability. The classifier uses spatio (from autoencoder embedding) and temporal context information (Bi-GRU) for the classification task. An attention mechanism is also incorporated in the classifier module for better temporal modeling. The classification results signify the potential of excitation source (ILPR) information in the current task. A significant performance improvement is observed for 1 sec segment duration in comparison to 500 ms. Fundamental frequency (F_0) plays an important role in shouted speech production. The F_0 also varies according to gender. Hence, this work can be further extended by studying the effect of gender on shouted speech detection.

6. References

[1] S. Devi, "Making sense of views culture in television news media in india," *Journalism Practice*, vol. 13, no. 9, pp. 1075–1090, 2019.

[2] J. Pohjalainen, T. Raitio, S. Yrttiaho, and P. Alku, "Detection of shouted speech in noise: Human and machine," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2377–2389, 2013.

[3] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream detection for home applications," in *IEEE Conference on Industrial Electronics and Applications*, June 2010, pp. 2115–2120.

[4] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2061–2071, 2009.

[5] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3050–3061, 2013.

[6] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, and P. Alku, "Analysis and synthesis of shouted speech," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013, pp. 1544–1548.

[7] P. Alku, T. Bckstrm, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *The Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[8] S. Baghel, S. R. Mahadeva Prasanna, and P. Guha, "Excitation source feature for discriminating shouted and normal speech," in *International Conference on Signal Processing and Communications (SPCOM)*, July 2018, pp. 167–171.

[9] S. Baghel, S. R. M. Prasanna, and P. Guha, "Analysis of excitation source characteristics for shouted and normal speech classification," in *2020 National Conference on Communications (NCC)*, 2020, pp. 1–6.

[10] S. Baghel, S. R. M. Prasanna, and P. Guha, "Exploration of excitation source information for shouted and normal speech classification," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 1250–1261, 2020.

[11] J. Pohjalainen, P. Alku, and T. Kinnunen, "Shout detection in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2011, pp. 4968–4971.

[12] H. Nanjo, T. Nishiura, and H. Kawano, "Acoustic-based security system: Towards robust understanding of emergency shout," in *International Conference on Information Assurance and Security*, vol. 1, Aug 2009, pp. 725–728.

[13] W. Liao and Y. Lin, "Classification of non-speech human sounds: Feature selection and snoring sound analysis," in *IEEE International Conference on Systems, Man and Cybernetics*, Oct 2009, pp. 2695–2700.

[14] V. K. Mittal and A. K. Vuppala, "Changes in shout features in automatically detected vowel regions," in *International Conference on Signal Processing and Communications*, June 2016, pp. 1–5.

[15] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732–742, 2012.

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

[17] S. Baghel, S. R. M. Prasanna, and P. Guha, "Classification of multi speaker shouted speech and single speaker normal speech," in *IEEE Region 10 Conference*, Nov 2017, pp. 2388–2392.

[18] S. Baghel, M. Bhattacharjee, S. R. M. Prasanna, and P. Guha, "Shouted and normal speech classification using 1d cnn," in *Pattern Recognition and Machine Intelligence*, B. Deka, P. Maji, S. Mitra, D. K. Bhattacharyya, P. K. Bora, and S. K. Pal, Eds. Cham: Springer International Publishing, 2019, pp. 472–480.

[19] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6460–6464.

[20] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2471–2480, Dec 2013.