Shouted and Normal Speech Classification Using 1D CNN

Shikha Baghel[1](✉), Mrinmoy Bhattacharjee[1], S. R. M. Prasanna[1,2], and Prithwijit Guha[1]

[1] Department of Electronics and Electrical Engineering,
Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India
{shikha.baghel,mrinmoy.bhattacharjee,prasanna,pguha}@iitg.ac.in
[2] Department of Electrical Engineering, Indian Institute of Technology Dharwad,
Dharwad 580011, India

Abstract. Automatic shouted speech detection systems usually model its spectral characteristics to differentiate it from normal speech. Mostly hand-crafted features have been explored for shouted speech detection. However, many works on audio processing suggest that approaches based on automatic feature learning are more robust than hand-crafted feature engineering. This work re-demonstrates this notion by proposing a $1D$-CNN architecture for shouted and normal speech classification task. The CNN learns features from the magnitude spectrum of speech frames. Classification is performed by fully connected layers at later stages of the network. Performance of the proposed architecture is evaluated on three datasets and validated against three existing approaches. As an additional contribution, a discussion of features learned by the CNN kernels is provided with relevant visualizations.

Keywords: Shouted and normal speech classification · Shouted speech detection · $1D$ CNN · Convolution filter visualization

1 Introduction

Automatic shouted speech detection has application in areas like health-care, security and home-care [12]. Moreover, it is also required as a preprocessing step in applications like ASR and speaker recognition systems. Performance of such systems that are mostly trained on normally phonated speech degrades when test utterances include shouted content [12,14]. Thus, this work focusses on efficient segregation of normal and shouted speech.

Production of shouted speech may be attributed to following situations – charged emotions while speaking, communicating over long-distances, or calling in distress. In such situations, speech production characteristics deviate from normal. The effect of change in production characteristics is reflected in vocal tract characteristics. Air pressure from lungs increases during shouting, leading to comparatively fast vibration of vocal folds [9]. This changes the fundamental

B. Deka et al. (Eds.): PReMI 2019, LNCS 11942, pp. 472–480, 2019.
https://doi.org/10.1007/978-3-030-34872-4_52

frequency (F_0) of produced speech. Properties of F_0 has been extensively studied in the context of shouted and normal speech. In literature, shouted speech is considered as a high vocal effort speech. Authors in [9] have studied deviation in vocal efforts for five vocal modes – whisper, soft, neutral (normal), loud and shout. Spectral tilt and log-linear predictive coding were examined in [14] to study the effect of different vocal efforts on automatic speech recognition. The classification performances were tested using three variants – (a) Bayesian classifier based on static GMMs with diagonal covariance matrices; (b) GMMs with full covariance matrices; (c) a multi-class Support Vector Machine (SVM) classifier with Radial Basis Function (RBF) kernel. To capture the variability in vocal tract characteristics due to different vocal modes, Mel-Frequency Cepstral Coefficient (MFCC) has been largely studied along with different combination of features [12]. Mittal et al. in [9] used features derived from the Hilbert envelope of double differenced Numerator of the Group Delay (HNGD) spectrum to analyze the differences between shouted and normal speech. Strength of Excitation (SoE), F_0 and dominant frequency (F_D) derived from the LP spectrum of vowel-like regions were examined to discriminate shouted and normal speech. Shifts in formant positions are also used in literature for detection of shouted speech. Recently, the standard deviation in frequency and energy of the first three formants have been explored for analyzing and characterizing shouted speech signals [8].

Existing works have mostly focused on using excitation or spectral features for detection of shouted speech [10]. Recent works in deep learning have shown that effective architectures can learn robust and discriminative patterns from data and provide promising classification performance [2]. Such representations are learned with a framework of many affine transformations followed by a nonlinearity [4]. These automatically learned features are comparatively more robust to local variations in data. Convolutional Neural Networks (CNN, henceforth) have been used for feature learning from audio data for several audio processing applications like speech-based emotion recognition [3], speech recognition [11], hate speech detection [1] etc. CNNs are capable of capturing patterns present in data. This motivated us to explore CNNs for shouted and normal speech classification. Motivated by the success of CNNs in various image processing applications [7], authors have deployed $2D$ CNNs on spectrograms (or other time-frequency variants) of audio data [2,3]. However, a drawback of this approach is the requirement of large datasets for learning the networks. To the best of our knowledge, most standard datasets (available in public domain) for shouted speech classification are small in size. Other works have reported results on private datasets (remains undisclosed). Hence, it is difficult to get a sufficiently large dataset for employing $2D$-CNNs for shouted and normal speech classification. This motivated us to explore $1D$-CNNs consisting of one-dimensional convolutional kernels. We propose a $1D$-CNN with two convolution layers, one max-pooling stage and three dense layers (Fig. 1). The proposed CNN is benchmarked on three standard datasets with respect to three baseline algorithms. To summarize, this work has the following contributions. First, a proposal of a

1D CNN architecture for shouted and normal speech classification. Additionally, CNN filters are analyzed in an attempt to interpret the learned features. Second, construction of a dataset of 20 speakers in both shouted and normal vocal mode.

The rest of the paper is organized in the following manner. The proposed approach is described in Sect. 2. The experimental results on standard datasets are presented and discussed in Sect. 3. Learned filters are visualized in Sect. 4. We conclude in Sect. 5 and sketch the future scope of present work.

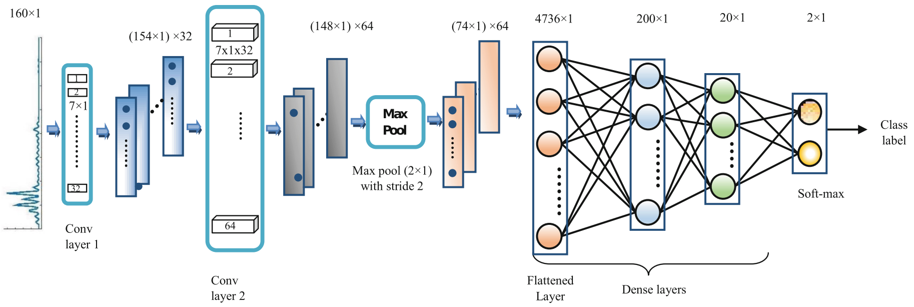

Fig. 1. Proposed 1D CNN architecture.

2 Proposed Approach

Existing works in shouted and normal speech classification have mostly used spectral features, like spectral tilt [14], formant locations [8], MFCCs [14] etc. This motivated us to first transform the input speech signal to frequency domain. Due to quasi-periodic nature of speech, it is segmented in small overlapping frames. Each frame $s[n]$ $(n = 0, 1, \ldots 2N - 1)$ of length $2N$ is transformed to frequency domain as $S[k] = \sum_{n=0}^{2N-1} s[n]e^{\frac{-j2\pi kn}{2N}}; k = 0, 1, \ldots, 2N-1$, where $S[k]$ represents spectrum of $s[n]$. Since, $s[n]$ is real, only half the magnitude spectrum $(S[k];\ k = 0, 1, \ldots N - 1)$ is used as input to 1D CNN. This work uses a frame size of 20 ms with a shift of 10 ms. We have used audio signals sampled at 16 KHz. Thus, each frame of 20 ms interval contains $2N = 320$ samples. Hence, input to the 1D CNN has $N = 160$ dimensions. The proposed architecture of the 1D CNN is described next.

A conventional architecture of 1D CNN comprises sets of convolutional and pooling layers followed by fully connected dense layers. Small-sized 1D filters are used in each convolutional layer to capture the local feature of input data. The output of convolutions is passed through an activation function for non-linear transformation of data. Pooling is performed to obtain the hierarchical representation in data. All tunable parameters of the architecture are learned through a feed-forward and back-propagation approach, which minimizes a cost function. The rest of the paper is organized in the following manner. The proposed approach is described in Sect. 2. The experimental results on standard datasets are presented and discussed in Sect. 3. Learned filters are visualized in Sect. 4. We conclude in Sect. 5 and sketch the future scope of present work. The $1D$ CNN is illustrated in Fig. 1.

Table 1. Classification performance in terms of F-score

	SNE-Speech		FIN-SN		IIIT-H VLSD	
	$\mathrm{F_{nor}}$	$\mathrm{F_{sh}}$	$\mathrm{F_{nor}}$	$\mathrm{F_{sh}}$	$\mathrm{F_{nor}}$	$\mathrm{F_{sh}}$
Raitio-FS [13]	0.79 ± 0.01	0.79 ± 0.00	0.9 ± 0.01	0.89 ± 0.00	0.72 ± 0.01	0.74 ± 0.02
Mittal-FS [9]	0.83 ± 0.01	0.79 ± 0.01	0.92 ± 0.01	0.90 ± 0.00	0.76 ± 0.02	0.79 ± 0.02
Zelinka-FS [14]	0.95 ± 0.01	0.93 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.87 ± 0.01	0.86 ± 0.01
$1D$-CNN	0.97 ± 0.003	0.96 ± 0.003	0.99 ± 0.002	0.99 ± 0.001	0.9 ± 0.01	0.9 ± 0.01

Convolutional Layer. Convolutional operations are performed on local regions of input data through different kernels. Each convolutional kernel extracts certain patterns from the input data. The mathematical representation of convolutional operation in layer l is given by $\mathbf{y}_i^l = f\left(\mathbf{x}^l \star \mathbf{K}_i^l + b_i^l\right)$ where, $\mathbf{K}$ denotes the convolutional kernel, i corresponds to kernel index and b_i^l is the bias of i^{th} kernel in l^{th} layer. Here, $\mathbf{x}^l$ and $\mathbf{y}_i^l$ represent the input and output of l^{th} layer respectively. The convolution operator is represented by $\star$ and $f(.)$ corresponds to activation function. The proposed architecture has two convolutional layers with 32 and 64 kernels respectively. Kernel size in both the layers is 7×1. ReLU $(max(0, x))$ is used as an activation function to address the vanishing gradient problem [6]. The 160×1 input is processed by these two convolutional layers to produce a $(148 \times 1) \times 64$ tensor.

Max-Pooling. Pooling operation is used to downsample data for reducing model complexity of CNN. Pooling retains necessary representational information for further layers. A single 2×1 max-pooling stage (with stride 2) is used in our proposal. This max-pooling stage downsamples the $(148 \times 1) \times 64$ tensor to $(74 \times 1) \times 64$.

Dense Layers. Fully connected dense layers are used at the later stages of the CNN. Each node in such a layer is connected to all nodes in the previous layer. The output of a dense layer node is mathematically represented as $\mathbf{d}^l = f(\mathbf{W}^l \mathbf{d}^{l-1} + \mathbf{b}^l)$, where, $\mathbf{W}$ and $\mathbf{b}$ represent weight and bias respectively. The $(74 \times 1) \times 64$ tensor is flattened to obtain a layer containing 4736 nodes. This is followed by two hidden layers with 200 and 20 nodes respectively till last (output) layer. Each dense layer uses a ReLU activation function. The last layer hosts soft-max activation function and has 2 nodes corresponding to shouted and normal speech.

Model Training. The predicted and actual output are used to estimate the prediction error of the CNN. The CNN parameters are optimized using cross-entropy loss function with ADAM optimizer [5]. Parameters of the model are optimized by minimizing the error with a learning rate of 0.0001. We have trained the CNN for a maximum number of 150 epochs with a mini-batch size of 512.

Table 2. Cross-dataset classification performance in terms of F-score for $1D$-CNN.

	SNE-Speech		FIN-SN		IIIT-H VLSD	
	$\mathrm{F_{nor}}$	$\mathrm{F_{sh}}$	$\mathrm{F_{nor}}$	$\mathrm{F_{sh}}$	$\mathrm{F_{nor}}$	$\mathrm{F_{sh}}$
SNE-Speech	–	–	0.54 ± 0.044	0.78 ± 0.01	0.65 ± 0.03	0.73 ± 0.002
FIN-SN [12]	0.86 ± 0.01	0.82 ± 0.01	–	–	0.82 ± 0.01	0.78 ± 0.01
IIIT-H VLSD [9]	0.80 ± 0.02	0.80 ± 0.02	0.79 ± 0.03	0.86 ± 0.01	–	–

3 Experiments and Results

The proposed approach is evaluated on three datasets. The first dataset (FIN-SN, henceforth) contains 1024 Finnish sentences uttered by 22 (11 males and 11 females) native speakers of Finnish language [12][1]. The speech signal was recorded at 16 kHz for both normal and shouted vocal modes. Second, the IIIT-H Volume Level Study Database (IIIT-H VLSD, henceforth) [9] (See footnote 1) comprises normal and shouted data of 17 (10 males and 7 females) non-native English speakers. A total of 102 English sentences are recorded at 48 kHz sampling rate. Third, a dataset contributed by the authors (SNE-Speech, henceforth) consists of 20 (10 female and 10 male) non-native English speakers. Speakers were asked to utter 30 English sentences in normal and shouted vocal modes. Speakers belong to different states of India. Therefore, the accent may vary from speaker to speaker. Out of 30 sentences, 15 are in imperative mood that people might use to threaten someone. While remaining 15 sentences have a neutral mood. All recordings were done in a controlled environment with TASCAM DR-100mkII 2-channel portable digital recorder, electroglottograph (with two electrodes) and Praat software. Speech and corresponding Electroglottogram (EGG) signals were sampled at 44.1 kHz. The dataset comprises a total of 1200 sentences in both normal and shouted vocal mode. All audio signals of all three datasets are resampled at 16 kHz for our experiments.

Our proposal is validated against three baseline approaches. Features used to analyze shouted speech in [13] are considered as baseline Raitio-FS. These features are F_0, Normalized Amplitude Quotient (NAQ), Sound Pressure Level (SPL) and the difference between first and second harmonics (H1-H2). The effect of varying the glottal dynamics on speech production due to different vocal modes are analyzed in [9]. The F_0, alpha (ratio of closed phase to the glottal cycle), beta (ratio of low-frequency energy to the high-frequency energy in normalized HNGD spectrum) and standard-deviation of low-frequency energy have been reported to have different characteristics in different vocal modes. This

[1] Authors of this work would like to thank Pohjalainen et al. [12] and Mittal et al. [9] for sharing their dataset.

work is considered as baseline Mittal-FS [9]. Baseline Zelinka-FS [14] reported classification of five vocal modes using 20-MFCCs using 40 filters. Authors in [9, 13] presented an analysis of shouted and normal speech. They did not report any classification result. For the comparison purpose, classification is done using SVM classifier. Classification performance for shouted (F_{sh}) and normal (F_{nor}) speech are reported in terms of F1-scores of individual classes.

We have used SVM (with RBF kernel) for the classification task of baseline approaches. Grid search is performed to find the optimal classifier parameters. All experiments are carried out with train-test split of 80:20. For each approach, the classification task is performed five times on randomly drawn instances of training and testing. Each reported result in Table 1 represents the mean and standard-deviation of F-scores for all five experiments. Classification performance of 20-MFCCs is higher than the other two baselines in all three datasets. The MFCC gives almost similar performance as of $1D$ CNN for FIN-SN dataset. The $1D$ CNN provides comparatively better classification results in all three datasets. All the approaches show lower results in IIIT-H VLSD dataset. This dataset contains few shouted recordings which are perceptually similar to normal speech. This may be the reason for lower performance in this dataset. Classification results using $1D$-CNN validate its ability of efficient representation learning.

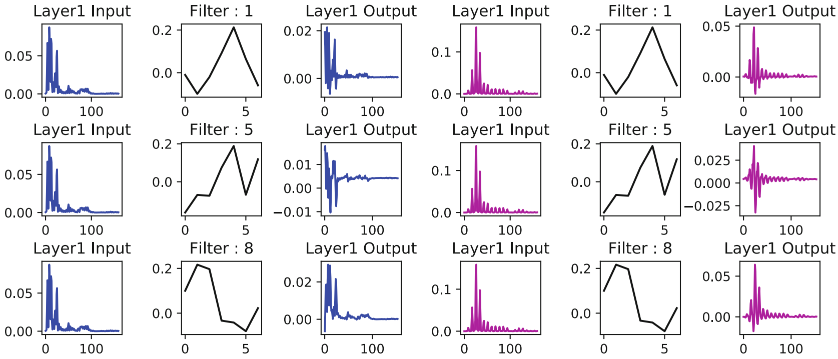

Fig. 2. Visualization of responses of few selected filters from the first convolution layer of the proposed CNN architecture as shown in Fig. 1. **First** and **fourth** columns correspond to inputs to the filters for normal and shouted speech respectively. Similarly, **third** and **sixth** columns correspond to outputs of the filters for normal and shouted speech respectively.

4 Analysis and Filter Visualization

The convolution filters of CNNs learn some specific properties of input data. The proposed $1D$-CNN is trained on the DFT spectrum of speech frames. Thus, filters of the proposed CNN might learn spectral characteristics of normal and shouted speech. We attempt to develop some intuitive understanding of the features learned by these filters. Figure 2 shows the shapes of some selected filters from the first convolution layer (CL, henceforth) of the trained CNN. We observe that each filter in first CL learns highly varying harmonic content of the input spectrum. The learned harmonic patterns seem to be quite different for normal and shouted speech (Fig. 2). We believe that this representation learning leads to better discrimination between the two classes.

A cross-dataset performance analysis is performed to establish the generalization performance of our proposal. The CNN is trained on one dataset and tested on the other two datasets. Table 2 shows the cross dataset performance. The results indicate the dependency of the CNN model on the training data up to some extent. Datasets FIN-SN and SNE-Speech have different languages. Dataset IIIT-H VLSD has a comparatively smaller corpus and accent of speakers is quite different from that of SNE-Speech. These might be the reasons of lower classification performance for cross datasets. This motivated us to train the CNN on combined dataset.

Table 3. Classification performance in terms of F-score for combined dataset.

Training data	$1D$-CNN	
	F_{nor}	F_{sh}
40% of SNE-Speech + 40% of FIN-SN + 40% of IIIT-H VLSD	0.96 ± 0.005	0.97 ± 0.009
60% of SNE-Speech	0.95 ± 0.012	0.96 ± 0.006
60% of FIN-SN	0.98 ± 0.003	0.99 ± 0.004
60% of IIIT-H VLSD	0.86 ± 0.029	0.85 ± 0.025

Proposed CNN architecture is trained on combined data drawn from all three datasets to make it robust towards language and accent variations. For training, 60% of all three datasets (60% of SNE-Speech + 60% of FIN-SN + 60% of IIIT-H VLSD) are randomly drawn and remaining data is used for testing. Table 3 illustrates the performance of the $1D$ CNN model trained on combined data. The results show that the trained model gives almost similar performance as mentioned in Table 1 except for IIIT-H VLSD dataset. IIIT-H VLSD is comparatively smaller dataset than the other two. Thus, the trained model might be biased towards the other two datasets. This conveys that the trained model is relatively robust towards language and accent than the one learned from individual datasets.

5 Conclusion and Future Scope

This work proposed a $1D$ CNN architecture for shouted and normal speech classification. The proposal is evaluated on three datasets and validated against three baseline methods. The CNN was able to efficiently learn discriminating features from the magnitude spectrum of speech frames. CNN convolution layer visualization reveals that filters in the first convolution layer learn the harmonic structure of speech spectra. Efficacy of the learned features is validated by decent generalization performance obtained with cross-data and combined data performance analysis. The CNN performance indicates that the proposed approach is effective for shouted and normal speech classification.

In the future, further exploration of the effect of different kernel and convolution layer sizes can be explored. Slight degradation in performance of cross-dataset experiments indicates the dependency of the model on language. Thus, further explorations are required to design language-independent CNN based methods. The performance of the proposed CNN based approach for different noisy speech conditions also requires further investigations.

References

1. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion, WWW 2017 Companion, pp. 759–760. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva (2017)
2. Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W.: Speech emotion recognition from spectrograms with deep convolutional neural network. In: Proceedings of 2017 International Conference on Platform Technology and Service (PlatCon), pp. 1–5, February 2017
3. Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM International Conference on Multimedia, MM 2014, pp. 801–804. ACM, New York (2014)
4. Kim, J., Englebienne, G., Truong, K.P., Evers, V.: Deep temporal models using identity skip-connections for speech emotion recognition. In: Proceedings of the 25th ACM International Conference on Multimedia, MM 2017, pp. 1006–1013. ACM, New York (2017)
5. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
7. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
8. Mesbahi, L., Sodoyer, D., Ambellouis, S.: Shout analysis and characterisation. Int. J. Speech Technol. **22**(2), 295–304 (2019)
9. Mittal, V.K., Yegnanarayana, B.: Effect of glottal dynamics in the production of shouted speech. J. Acoust. Soc. Am. **133**(5), 3050–3061 (2013)
10. Mittal, V.K., Yegnanarayana, B.: An automatic shout detection system using speech production features. In: Böck, R., Bonin, F., Campbell, N., Poppe, R. (eds.) MA3HMI 2014. LNCS (LNAI), vol. 8757, pp. 88–98. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15557-9_9
11. Palaz, D., Magimai-Doss, M., Collobert, R.: Analysis of CNN-based speech recognition system using raw speech as input. In: Proceedings of Sixteenth Annual Conference of the International Speech Communication Association, pp. 11–15 (2015)

12. Pohjalainen, J., Raitio, T., Yrttiaho, S., Alku, P.: Detection of shouted speech in noise: human and machine. J. Acoust. Soc. Am. **133**(4), 2377–2389 (2013)
13. Raitio, T., Suni, A., Pohjalainen, J., Airaksinen, M., Vainio, M., Alku, P.: Analysis and synthesis of shouted speech. In: Proceedings of Fourteenth Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1544–1548 (2013)
14. Zelinka, P., Sigmund, M., Schimmel, J.: Impact of vocal effort variability on automatic speech recognition. Speech Commun. **54**(6), 732–742 (2012)