Determining Number of Speakers from Single Microphone Speech Signals by Multi-Label Convolutional Neural Network

Haoran Wei
Department of Electrical & Computer Engineering
University of Texas at Dallas
Richardson, TX 75080, USA
Haoran.Wei@utdallas.edu

Nasser Kehtarnavaz
Department of Electrical & Computer Engineering
University of Texas at Dallas
Richardson, TX 75080, USA
kehtar@utdallas.edu

***Abstract* — This paper presents a multi-label convolutional neural network approach to determine the number of speakers when using a single microphone which is more challenging than when using multiple microphones. Spectrograms of windowed noisy speech signals for 1talker, 2talkers and 3+talkers are used as inputs to a multi-label convolutional neural network. The architecture of the developed multi-label convolutional neural network is discussed and it is shown that this network with median filtering can achieve an overall accuracy of about 81% for the noisy speech dataset examined.**

Keywords —determining number of speakers, multi-label convolutional neural network, speech processing

I. Introduction

Determining the number of speakers in multi-speaker audio environments is used in many speech processing applications such as speaker tracking [1, 2], speaker clustering [3], speaker diarization [4], and speaker identification [5, 6]. Multiple microphones are normally considered for determining the number of speakers, e.g. [7-9], where the differences of the time delays of speech signals from different microphones are used to determine the number of speakers. Considering that a single microphone is used in many devices, it is more challenging to determine the number of speakers based on the signal from only a single microphone. This paper addresses this more challenging scenario.

In [10], the number of speakers was found by examining the modulation index between 2Hz and 8Hz. This index decreased as the number of speakers was increased. In this work, only eight sentences were examined and no accuracy outcome was reported. In [11, 12], a confidence parameter based on the statistical characteristics of the Mel filter was used to obtain the number of speakers. The results reported in these works exhibited precision when the number of speakers was less than four, however, the presence of noise in audio environments was not taken into consideration.

Basically, in the previous works, the focus has been placed on identifying effective features and no comprehensive dataset has been examined. This paper differs from the previous works in three ways: (1) The availability of speech signal is limited to a single microphone. (2) A convolutional neural network (CNN) is deployed to determine the number of speakers. As a result, feature selection is done by the network itself. (3) A more comprehensive speech dataset is considered in the presence of noise compared to the limited datasets used in the previous works.

The rest of the paper is organized as follows: a description of the dataset and the spectrogram used as the input to a multi-label CNN is provided in Section II. The architecture of the developed multi-label CNN is covered in Section III. The experimental results and their discussion are reported in Section IV. Finally, the conclusion is stated in Section V.

II. Dataset and CNN Input Considered

The dataset used in this study consists of speech signals collected by the Beijing Haitian Ruisheng Science Technology Limited [13]. These signals consists of 12349 utterances in English spoken by 40 speakers (20 from the US and 20 from the UK) with 20 of them being females and 20 of them males with ages ranging from 17 to 55 years old. The audio of 10 of the speakers were recorded by an iPhone, 10 by an Android smartphone, and 20 by a desktop computer.

The above dataset was randomly divided into 32 speakers (16 females and 16 males) for training and 8 speakers (4 females and 4 males) for testing. Then, their audio signals were mixed to generate 2talkers, and 3+talkers (babble) audio streams. Fig. 1 provides an illustration of the mixing performed to generate multi-talker signals. At the same time, machinery noise was added to all the signals at two SNR levels of 5dB and 10dB. As a result, the 2talkers data consist of 32 noisy audio streams of two talkers and the 3+talkers data consist of 32 noisy audio streams of three as well as more than three talkers. Table I shows a listing of

978-1-5090-6684-1/18/$31.00 ©2018 IEEE

the dataset audio streams.

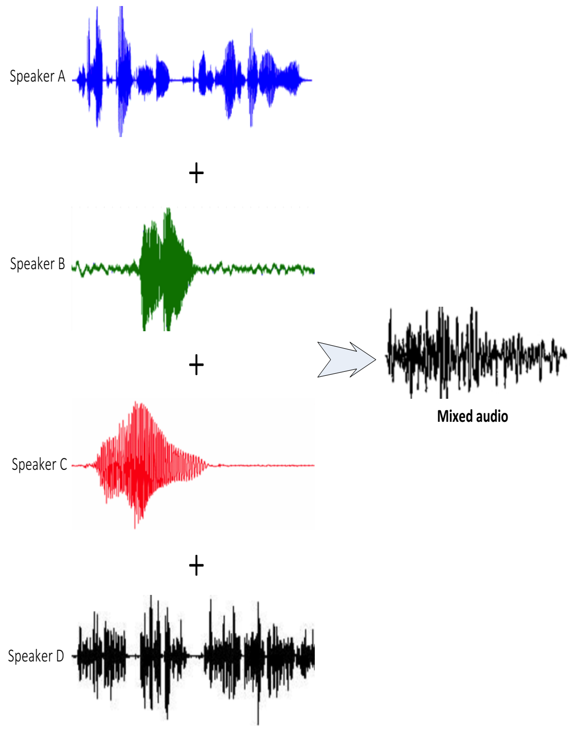

Fig. 1. Mixing of speech signals to generate multi-talker audio streams

TABLE I. DATASET EXAMINED

	Training set			Testing set		
SNR	Clean	5dB	10dB	Clean	5dB	10dB
1 talker	16Desktop+12Android + 4iPhone captured audio streams			4Desktop+3Android + 1iPhone captured audio streams		
2 talkers	32 audio streams			8 audio streams		

3+ talkers	3talkers *8 audio streams;
	4talkers *8 audio streams;
	5talkers *4 audio streams;
	6talkers *4 audio streams;
	7talkers *4 audio streams;
	8talkers *4 audio streams;
				3talkers *2 audio streams;
				4talkers *2 audio streams;
				5talkers *1 audio stream;
				6talkers *1 audio stream;
				7talkers *1 audio stream;
				8talkers *1 audio stream;		

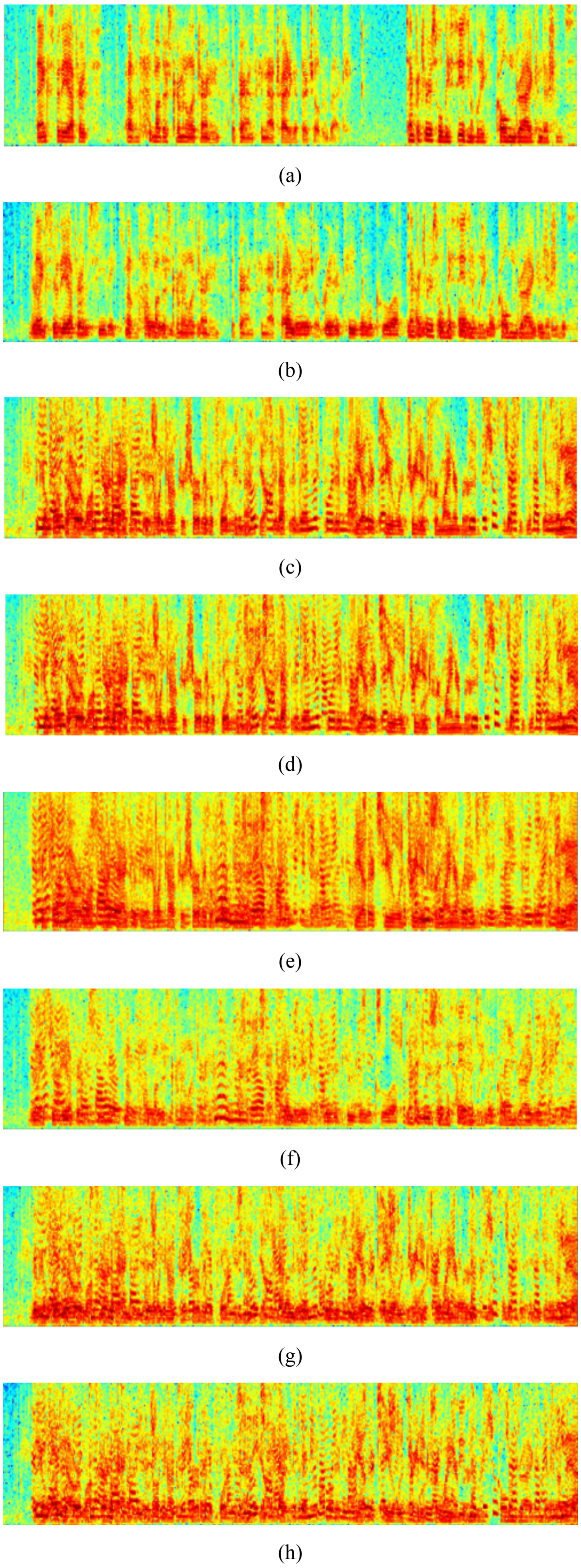
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

Fig. 2. Spectrograms of mixed speech signals from (a) one, (b) two, (c) three, (d) four, (e) five, (f) six, (g) seven, and (h) eight speakers or talkers; y-axis indicates frequency 0-4kHz & x-axis indicates time 0-10 sec.

From the audio streams, spectrograms are generated and represented as images. Spectrograms have been widely used in speech processing. A spectrogram image denotes intensity plots of the magnitude of the short-time Fourier transform (STFT) in log scale. The sampling rate used in these audio streams is 8kHz and the STFT frame size used is 25ms with 50% overlap between consecutive frames. This frame size is commonly used in speech processing. Fig. 2 illustrates sample spectrogram images for 1talker to 8talkers audio streams. These spectrograms images are then fed into a convolutional neural network discussed next.

III. Multi-Label Convolutional Neural Network Classification

Convolutional neural networks (CNNs) are being increasingly used in various classification and recognition tasks [14]. A CNN consists of numerous processing elements or units arranged in stacked layers comprising an input layer, a number of convolution and pooling layers, one or a few fully connected layers, and an output layer. After testing the performance of several different CNN architectures, the CNN architecture shown in Fig. 3 was found to be most effective for determining the number of speakers.

The input layer of the architecture shown in Fig. 3 corresponds to a spectrogram image. A typical convolution layer involves many filters. Here, three convolution layers are considered consisting of 16, 32 and 64 filters, respectively, all of size 3x3. Convolution layers capture local structures in spectrograms. Pooling layers merge the filtering operations of the convolution layers. The most common pooling is max pooling (winner takes all). The pooling size used here is 2x2. Batch normalization layers are considered for the purpose of normalizing data across a batch. These layers speed up the training and reduce the sensitivity to initialization. Rectified Linear Units (ReLUs) layers, which are the most commonly used activation function, follow the batch normalization layers. At the last stage, there is one or a few fully connected layers. A fully connected layer is similar to a layer of the well-known backpropagation neural network. The number of speakers is then determined from the output layer using the softmax loss function.

Normally, a single loss function is used for training. Performance is often improved when more than one loss function is considered. The architecture involving more than one loss function is named multi-label CNN, also named multi-task CNN. More details on multi-label CNN can be found in [15].

In the CNN architecture shown in Fig. 3, the network learns the differences between the three classes, that is 1talker, 2talkers and 3+talkers. To take into consideration differences between different 3+talkers situations, another CNN using the sigmoid loss function is used. The response of this sigmoid CNN is a value between 0.1 to 0.8 reflecting 1talker to 8talkers, respectively.

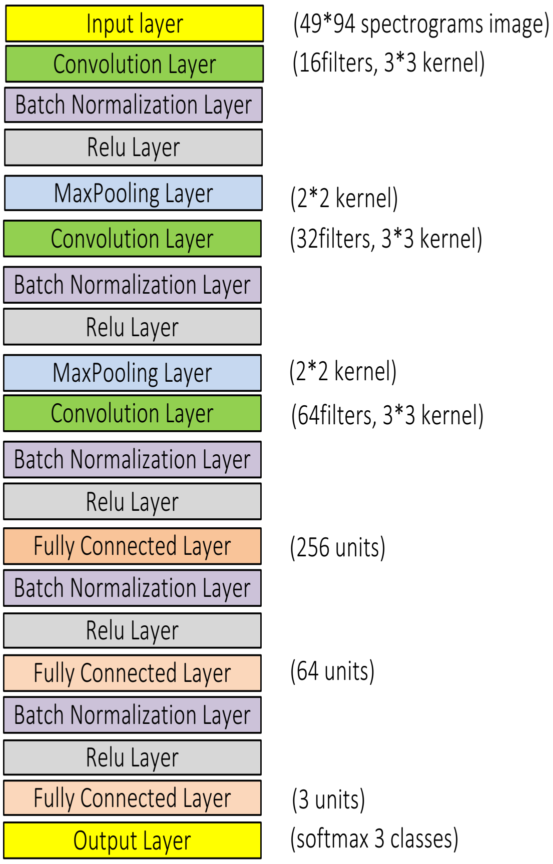

Fig. 3. Different layers of convolutional neural network

Next, the above two CNNs with different loss functions are trained separately. The details of the training process of a typical CNN can be found in [16]. The 256 outputs of the fully connected layer of the softmax CNN and the 64 outputs of the fully connected layer of the sigmoid CNN are combined together feeding another fully connected neural network to form our multi-label CNN as illustrated in Fig. 4. After conducting the multi-label CNN classification, a median filter of size 3 is applied to the output decision in order to smooth out fluctuations in the classification outcome.

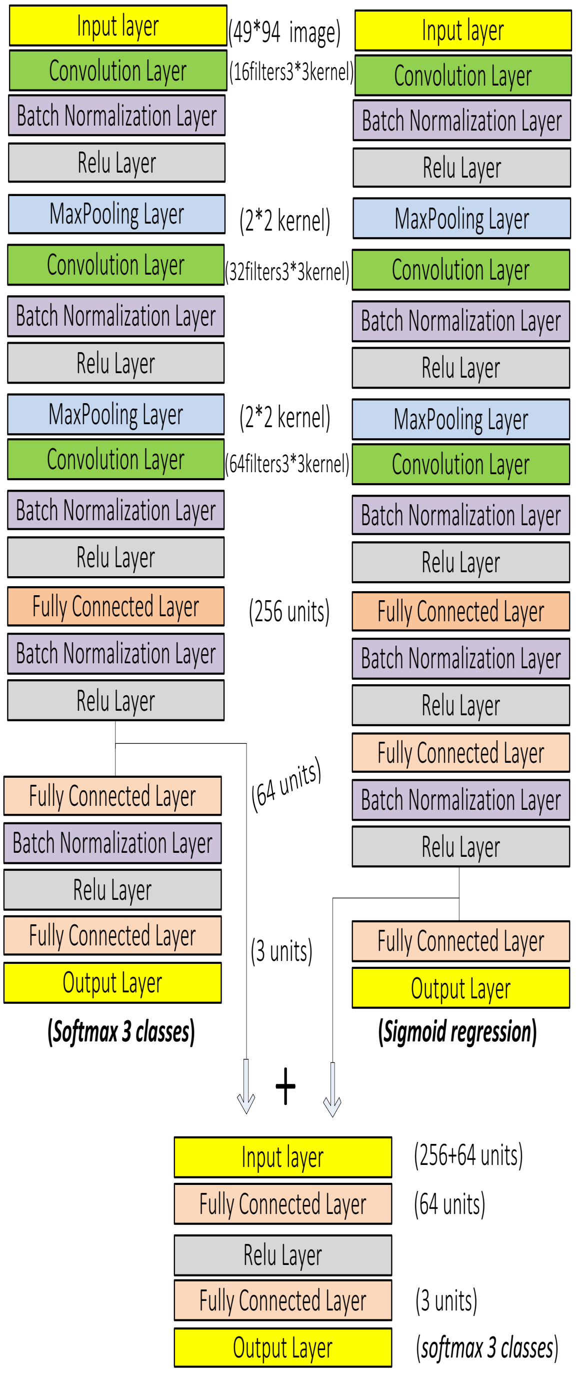

Fig. 4. Multi-label CNN architecture

IV. Experimental Results and Discussion

The performance of the developed approach was examined by comparing the class labels of the testing data obtained by the multi-label CNN with the ground truth class labels.

Table II lists the overall accuracies when using the following approaches: CNN, CNN with median filtering, multi-label CNN, and multi-label CNN with median filtering. The highest overall accuracy of about 81% was found when using the multi-label CNN with median filtering.

TABLE II. Comparison of approaches

Approach	Accuracy
CNN	72.6%
CNN + median filter	79.5%
Multi-label CNN	74.6%
Multi-label CNN + median filter	81.3%

The confusion matrix associated with the CNN with median filtering is shown in Table III and the confusion matrix associated with the multi-label CNN with median filtering is shown in Table IV. From these tables, it can be observed that the overlap between 2talkers and 3+talkers classes is the largest due to the similarities of their spectrograms.

TABLE III. Confusion matrix of CNN with median filtering

Identified class True class	1 talker	2 talkers	3+ talkers
1 talker	88.7%	11.1%	0.2%
2 talkers	11.2%	78.1%	10.7%
3+ talkers	0.3%	28.0%	71.7%

TABLE IV. Confusion matrix of multi-label CNN with median filtering

Identified class True class	1 talker	2 talkers	3+ talkers
1 talker	88.6%	11.3%	0.1%
2 talkers	10.1%	80.4%	9.5%
3+ talkers	0.2%	24.9%	74.9%

Another experiment was conducted to see how the multi-label CNN approach compared to a conventional classification approach. For this purpose, the Random Forest (RF) classifier which is found to be an effective classifier in a number of speech processing applications, e.g. [17], was used. The confusion matrix for this classifier is shown in Table V. This table shows that the overall accuracy when using the RF classifier was only about 60% which was far less than the accuracy when using the CNN classifier.

TABLE V. CONFUSION MATRIX OF RANDOM FOREST CLASSIFIER

Identified class True class	1 talker	2 talkers	3+ talkers
1 talker	63.6%	28.2%	8.2%
2 talkers	29.2%	42.7%	28.1%
3+ talkers	8.0%	24.4%	67.6%

V. Conclusion

A multi-label convolutional neural network for obtaining the number of speakers when using a single microphone has been developed in this paper. The use of a single microphone is more challenging than when multiple microphones are used due to the absence of time delays. A database of speech signals consisting of 1talker, 2talkers, and 3+talkers has been examined in the presence of noise at 5dB and 10dB SNR levels. This is the first time such a large dataset with noise is examined for determining the number of speakers based on a single microphone. Several convolutional neural network approaches were examined. The experimentations conducted have revealed that the developed multi-label convolutional neural network with median filtering outperformed the other approaches and achieved a classification accuracy of about 81%.

Acknowledgment

This work was sponsored in part by Cochlear Ltd. through a contract with University of Texas at Dallas.

References

[1] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *Proccedings of DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127-132, Lansdowne, 1998.

[2] M. Fallon and S. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1409-1415, 2012.

[3] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 411-416, Virgin Islands, 2003.

[4] H. Hung, Y. Huang, G. Friedland, and D. Perez , "Estimating the dominant person in multi-party conversations using speaker diarization strategies," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2197-2200, Las Vegas, 2008.

[5] H. Gish, M. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 873-876, New York, 1991.

[6] M. Hasan , M. Jamil , M. Rabbani , and M. Rahman, "Speaker identification using mel frequency cepstral coefficients," *Proceedings of 3rd International Conference on Electrical & Computer Engineering*, pp. 28-30, Taiwan, 2004.

[7] R. Swamy, K. Murty, B. Yegnanarayana, "Determining Number of Speakers From Multispeaker Speech Signals Using Excitation Source Information," *IEEE Signal Processing Letters*, vol. 14, no. 7, pp. 481-484, 2007.

[8] N. Namratha and R. Kumaraswamy, "Determining number of speakers in multi-speaker condition with additive noise," *Proceedings of National Conference on Electronics, Signals, Communication Optimization*, pp. 137-141, Karnataka, 2015.

[9] P. Kumar, L. Balakrishna, C. Prakash, S. Gangashetty, "Bessel features for estimating number of speakers from multispeaker speech signals," *Proceedings of IEEE 18th International Conference on Systems, Signals and Image Processing*, pp. 1-4, Sarajevo, 2011.

[10] T. Arai, "Estimating number of speakers by the modulation characteristics of speech," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2: II-197, Hong Kong, 2003.

[11] H. Sayoud and S. Ouamour, "Proposal of a new confidence parameter estimating the number of speakers-an experimental investigation," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, no. 2, pp. 101-109, 2010.

[12] S. Ouamour, M. Guerti, and H. Sayoud, "PENS: a confidence parameter estimating the number of speakers," *Proceedings of Second ISCA Workshop on Experimental Linguistics*, Athens, 2008.

[13] Beijing Haitian Ruisheng Science Technology Limited, http://kingline.speechocean.com

[14] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8599-8603, Vancouver, 2013.

[15] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv*, 2017.

[16] J. Bouvrie, "Notes on convolutional neural networks," *Neural Nets*, 2006.

[17] F. Saki, A. Sehgal, I. Panahi, and N. Kehtarnavaz, "Smartphone-based real-time classification of noise signals using subband features and random forest classifier," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2204-2208, Shanghai, 2016.