Quality Evaluation of Reverberant Speech Based on Deep Learning

Samia Abd El-Moneim
Electronics and Electrical Communications Engineering Department, Faculty of Electronic Engineering, Menouf 32951, Menoufia University.

Mahmoud Saied
Electronics and Electrical Communications Engineering Department, Faculty of Electronic Engineering, Menouf 32951, Menoufia University.

M. A. Nassar
Electronics and Electrical Communications Engineering Department, Faculty of Electronic Engineering, Menouf 32951, Menoufia University.

Moawad I. Dessouky
Electronics and Electrical Communications Engineering Department, Faculty of Electronic Engineering, Menouf 32951, Menoufia University.

N. Ismail
Electronics and Electrical Communications Engineering Department, Faculty of Electronic Engineering, Menouf 32951, Menoufia University.

Adel Saleeb
Electronics and Electrical Communications Engineering Department, Faculty of Electronic Engineering, Menouf 32951, Menoufia University.

Adel S. El-Fishawy
Electronics and Electrical Communications Engineering Department, Faculty of Electronic Engineering, Menouf 32951, Menoufia University.

Fathi E. Abd El-Samie
Electronics and Electrical Communications Engineering Department, Faculty of Electronic Engineering, Menouf 32951, Menoufia University.

Abstract— **This paper presents an efficient approach for classification of speech signals as reverberant or not. The reverberation is a severe effect encountered in closed room. So, it may affect subsequent processes and deteriorate speech processing system performance. The spectrograms are utilized as images generated from speech signals to be classified with deep convolutional neural networks. Spectrogram and MFCC are used as features to be classified with Long Short Term Recurrent Neural Network (LSTM RNN). Two models are presented and compared. Simulation results up to 100% classification accuracy are obtained. This can help in perform an initial step in any speech processing system that comprises quality level classification.**

1. Introduction

Speech is the mean of communication among individuals, it obtains sufficient data about the speaker identity, age, emotion, dialect, … etc., beside the carried information. There are several applications of speech processing such as coding, speech recognition, synthesis, and speaker recognition [1]. This application requires speech recordings, recording speech may be in closed rooms which will be subject to multi-reflections, which will be combined with the original signal in a phenomenon known as reverberations. The reverberant speech has poor quality than original one, so for speech processing application it is very important to perceive whether the speech is reverb speech or not?. Reverberation is an important issue that has to be considered in speech application, it characterized by an important parameter named as reverberation time [2],[3]. The longer the reverberation time, the severity of the reverberation on speech and the worse the quality of the recorded speech signal. The focus of this paper is to evaluate or cluster the input speech waveform by labeling the input speech to normal or reverb one [4]. This step is imperative prior to any speech based application, to decide if the recorded speech is normal or reverb, to take fast decision about processing or compensating reverberation of speech signal before it driven to the application as in Fig. 1.

Any clustering technique employs two mode training and testing. Each mode has two stage; extracting some distinguishing attributes and classification. In this paper the speech wave is used in its 2-D form, this can be implemented by obtaining speech spectrogram which is a good image representation of speech [5]. Deep learning techniques are used here for classification such as LSTM neural network and deep CNN [6-8]. One benefit of using deep CNN, it can make both features extraction and classification via some kernels which use convolution method to exclude some features called feature maps [8]. The paper is arranged as follows. Section 2 covers the reverberation effect, some ideologies of room acoustics and how to model the reverberation effect of the speech through a comb filter. Section 3 discusses the spectrogram representation of speech. some principals and conception of deep learning techniques are introduced in section 4. In section 5, the proposed speech quality evaluation approach is presented. Section 6 gives the simulation results and final conclusion is given in section 7.

2. Reverberation phenomena

Reverberation and echo are two distinct concepts, the echo is the reflected signal from the original speech signal when the original speech travel considerable distance. While the reverberation is the combination of the original speech with multiple reflection versions. This may be displayed as additional sound sources added to the system [9-12].

Reverberation' can be also defined as the persistence of sound in a space after a sound source has been stopped, it occurs in closed rooms. When the distance is small, such as in a room or theater, the speech will be reflected back to the source in less than one-tenth of a second [9].

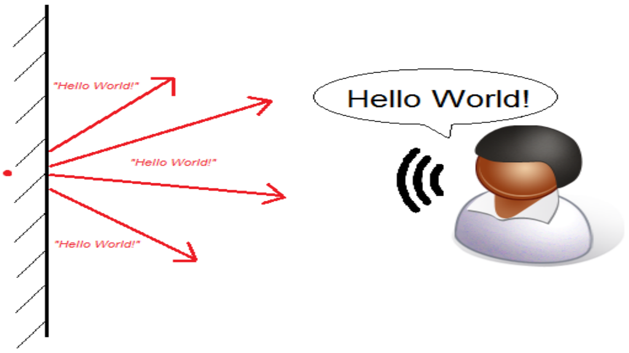


Fig. 2 Reverberation phenomena.

Only a small delay are in the speech replication, often only a few milliseconds, the listener perceived reverberation often as adding richness to the original speech. There are some associated parameter that characterize reverberation effect such as reverberation time which defined as the time in seconds taken for the sound to falloff by 60 dB from its initial value. High reverberation time can make a room sound loud and noisy which reduce the speech quality and intelligibility. Fig. 3 shows the influence of reverberation on the spectrogram of speech signal [10].

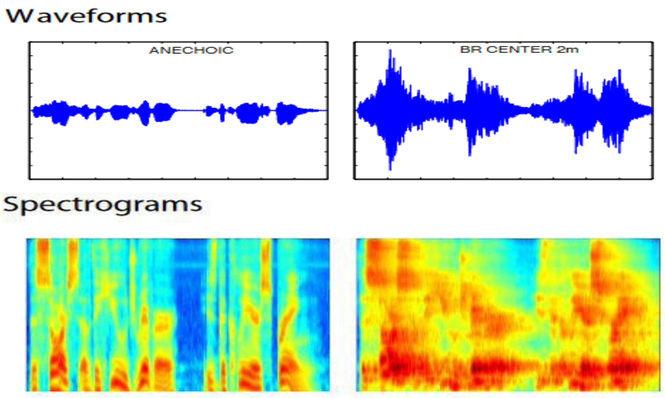


Fig. 3 The effect of reverberation on wave and spectrogram of speech.

The reverberant speech in closed room can be demonstrated as:

$$R(n) = s(n) * h(n) \quad [1]$$

Where h (n) is the room impulse response and s (n) is the original speech signal. The original speech is destroyed with long room impulse response. Reverberation of speech can be modeled by comb filter, it is implemented to add a delayed version of the signal to itself. If L is the filter length its impulse response is:

$$H(z) = 1 - z^{-L} \quad [2]$$

$$R(n) = s(n) - s(n-L) \quad [3]$$

Fig. 4 shows feed forward and feedback sections of comb filter simulating direct and reflected paths of speech signals. The degree of severity and quality of reverb speech can be determined by the reverberation time (RT60) as in Fig. 5 which can be defined as the time the speech signal takes to falling-off to 60 dB from its initial value [11],[12].

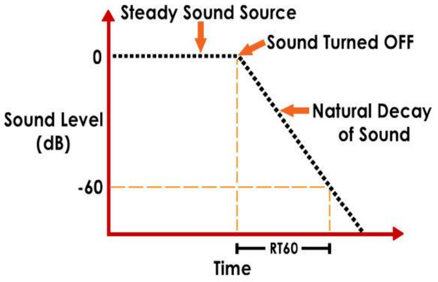


Fig. 5 Reverberation time.

3. Speech spectrogram

Spectrogram can represent the signal in the 2-D form, it include the complete information of an audio signal in both spectral and spatial domain as shown in Fig 6. A spectrogram is an image representation of the speech signal in the 2-D form. It can be calculated by taking the Short Time Fourier Transform (STFT) to the signal, by segmentation of the signal into segment of fixed length, then a window with a bit overlap is applied. The spectrogram is the squared magnitude of the STFT.

$$X(\tau,k) = \Psi\{x(n)\} = \sum_{n=0}^{N-1} x(n)w(n-\tau)e^{-jnk} \quad [4]$$

$$S(\tau,k) = |X(\tau,k)|^2 \quad [5]$$

Where Ψ is the STFT operator, X(τ,k) is the STFT of the signal x(n), w(n) is the window and S(τ,k) is the spectrogram [13], [14].

4. Deep learning

Deep learning is an extension of Neural Network (NN), since it is a NN with multi hidden layers. This makes them adept of demonstrating very composite and greatly nonlinear relations among inputs and output. Recurrent Neural Network (RNN) have been effectively applied to various sequence forecast and sequence classification tasks. Convolutional neural network (CNN) is a good image processing tools that can be used as feature extractor and classifier. So speech wave must be converted to image form to take the benefit of CNN to deal with speech.

4.1 Long Short Term Memory Recurrent Neural Network

Deep RNN has wide use in speech processing for its ability to label sequences, means that each input sequence is assigned to a certain class. RNN is a conventional neural network but with cyclic connections between its nodes. LSTM RNN is another version of RNN which replaces the hidden nodes by blocks. this blocks acts as a memory so it called memory blocks. The main mechanisms of an LSTM network are a sequence input layer and an LSTM layer. A sequence input [2] layer feedbacks sequence or time series data into the network, and the LSTM layer learns long-term reliance amongst time steps of sequence data. The below graph demonstrates the structure of a simple LSTM network for classification. The network begins with a sequence input layer followed by an LSTM layer. To forecast set labels, the network ends with a fully associated layer, a softmax layer and a classification layer [15].

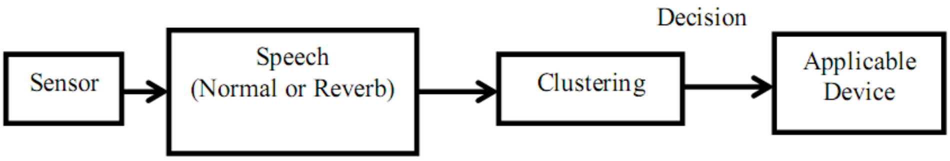


Fig. 1 Illustration of the speech clustering process.

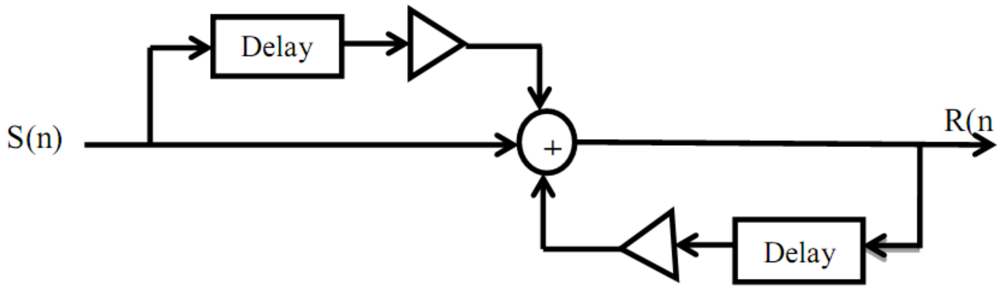


Fig. 4 Feedback and feedforward comb filter structure

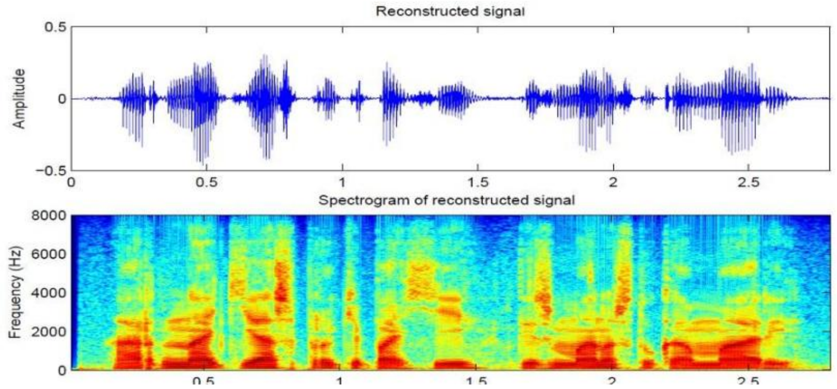


Fig. 6 Spectrogram of speech waveform.

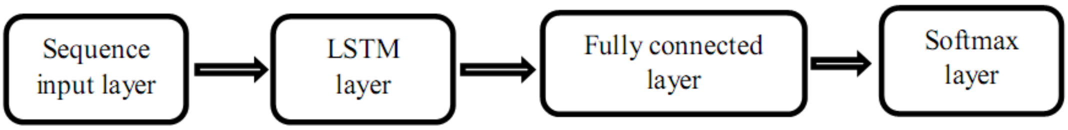


Fig. 7 Classification LSTM network

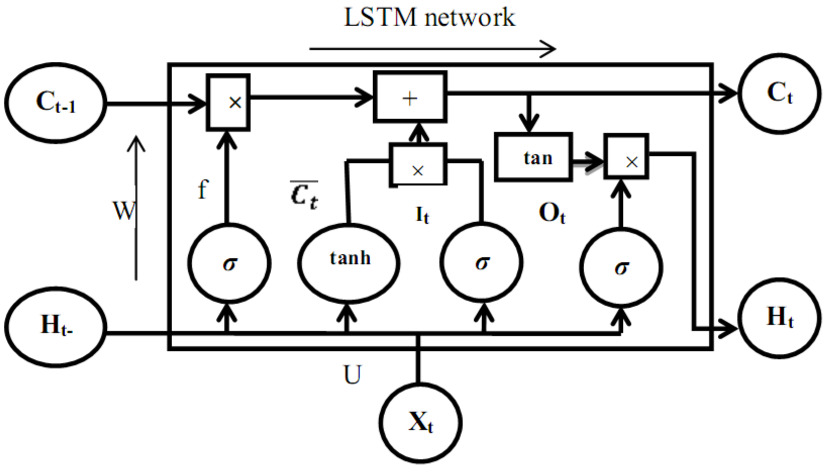


Fig. 8 LSTM memory block.

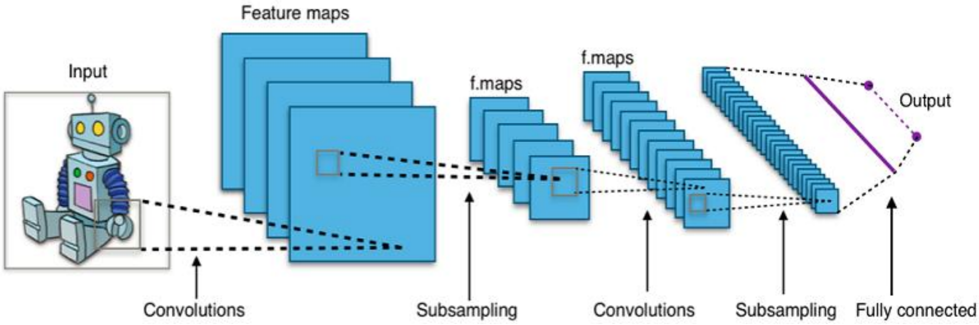


Fig. 9 CNN structure.

In every layer, the hidden nodes has replaced with blocks every block consist of four components named input gate, forget gate, cell and output gate. In Fig.8, The three gates control the state of the cell. At each time step, the layer either adds or removes information from the cell state, where the layer controls these updates using gates. Where the input gate control update of the cell state, forget the gate control reset of the cell state and output gate control the flow of output from the cell to another hidden layer [16].

Where X_t refers to the input , C_{t-1} is the previous cell memory, H_{t-1} is the previous cell output, C_t is the current cell memory and H_t the current cell output also U and W are weights. The mathematical model of the above architecture is illustrated by these equations.

$$f_t = \sigma\left(X_t * U_f + H_{t-1} * W_f\right) \quad [6]$$

$$\overline{C_t} = tanh(X_t * U_c + H_{t-1} * W_c) \quad [7]$$

$$I_t = \sigma(X_t * U_i + H_{t-1} * W_i) \quad [8]$$

$$O_t = \sigma(X_t * U_o + H_{t-1} * W_o) \quad [9]$$

$$C_t = f_t * C_{t-1} + I_t * \overline{C_t} \quad [10]$$

$$H_t = O_t * \tanh(C_t) \quad [11]$$

The memory blocks in the hidden layers of the LSTM turns as a memory that reserves the network's current state. The output of the softmax for a certain frame is a probability referring to one of the speakers as an output, but it based not only on the input frame but on every preexistent frame in this sequence [15], [16].

4.2 Convolutional Neural Network (CNN)

The CNN is a very good image processing tool, hence it can make both feature extraction an classification. In this paper the speech utterances were converted to its corresponding spectrogram image to benefit from CNN with speech. The CNN has two form of neural network, one for extracting features (convolutional layer CNV) from the spectrogram image and one more for classification as revealed in Fig. 9 which spell out the used system, so there is no prerequisite for feature extraction step, it extract robust features by itself [17].

The feature extraction neural network combines a pairs of layers convolutional layer (CNV) and pooling or subsampling layer. The CNV layer contains a group of digital filters named convolutional or kernel filters that uses the convolutional method to transform the input image into new image termed feature maps. The feature maps be different reliant on the used convolutional filters, these features is processed through an activation function before the output is yields. The pooling layer reduces the image dimension since it gathers the neighboring pixels into one pixel. The pooling pixel and the representative value is selected usually from square matrix with different number of pixels, then either yield the maximum or the mean value of the selected pixels. Fig. 10 is an illustration of the convolution, pooling and classification layer [17],[18].

5. Proposed speech clustering approach

The suggested scheme built on transforming the 1-D speech signal into its 2-D form, by obtaining its spectrogram as an image representation as in Fig 11. In this paper there are two cluster of speech: normal and reverb speech, reverberant speech are modeled by using comb filter and reverberation time (RT60 =0.5 sec). Any clustering techniques has two stage: training and testing, in training a model based on extracted features is made for each group and deposited in database, while in testing also a model is prepared and matched with that kept in database to find the greatest matches [5]. In LSTM spectrogram of speech is calculated and fed to the LSTM network as a 2-D feature vector, while in CNN the spectrogram is fed as an image and the network extract the features by itself and classification is achieved by obtainin the final decision through the fully connected layer.

6. Dataset description and results

The used data is a subset of much bigger data sets called Chinese Mandarin Corpus. This corpus were recorded in silence in-door environment using cellphone. Spectrogram is also obtained for each wave using STFT with window length of 256 sample. A 129 features was obtained by applying spectrogram, also 13 Mel Frequency Cepstrum Coefficient is extracted from speech corpus and used with LSTM net of input layer size that equal to the number of the input coefficients.. The whole feature vector enters the network at the same time, each coefficient corresponding to a node in the input layer. the network works as a sequence classification not frame classification, since the feature vectors from one speaker are seen as a sequence mapped to one target. A description of the used dataset is showing in table 1, also the training progress of LSTM and CNN are revealed in Fig. 12&13.. The accuracy measures the system performance which is defined by Equation 12 :

$$accuracy\ \% = \frac{Number\ \ of\ success\ \ clustering}{Total\ \ number\ \ of\ clustering\ \ trials} \times 100\% \quad [12]$$

Table 2 and Fig. 14 present the accuracy of the suggested method in case of using LSTM with both MFCC and spectrogram which reaches 100%, which demonstrates the ability of the suggested method to well distinguish between reverb and normal speech. Also it presents the accuracy when using CNN in case of two, three and four layer model which reaches 100% for the model of four layer.

Conclusion

This paper presented an efficient approach for quality level classification of speech signals. Two designed deep neural networks have been designed for this purpose. The performance of both deep neural networks have been compared. Simulation results proved that a 100% accuracy for the quality level classification approach can be achieved with 4-layer CNNs. Also, MFCC and spectrogram are used as features with LSTM and achieve accuracy of 100%.

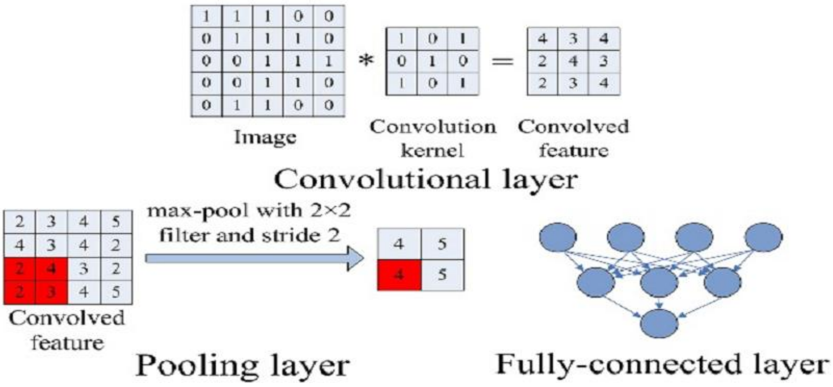


Fig. 10 Illustration of the convolution, the pooling and the fully-connected layer respectively.

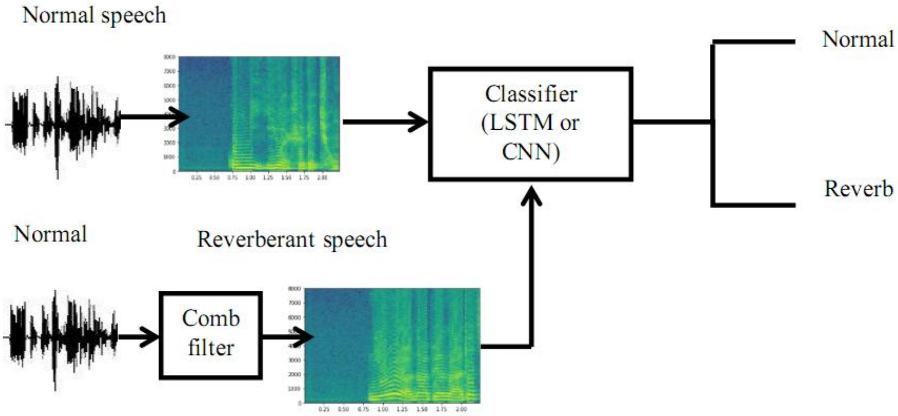
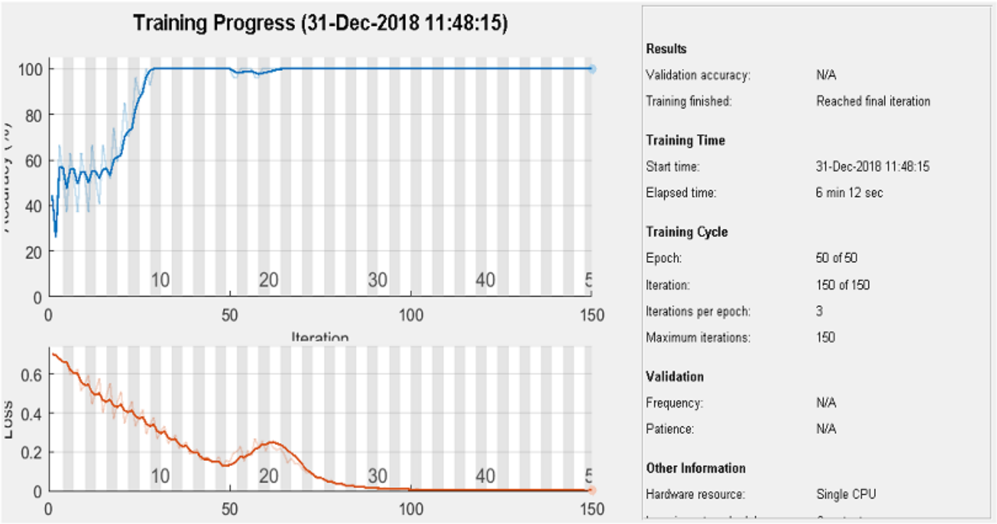


Fig. 11 Proposed speech clustering approach.

Table 1 Dataset description	
Number of training utterances	100 [50 normal and 50 with reverberation]
Number of testing utterances	30
Length and Sampling rate of speech	36000 samples, 16 KHZ respectively
Size of spectrogram image	224*224
Image type and format	RGB, png
Software applying algorithm	Matlab R2017b
Kernel filter size	3*3
Kernel number for layer 1, 2, 3 and 4	16, 32, 64 and 128 respectively
Pooling size	2*2
RT60	0.5 sec



Results

Validation accuracy: N/A

Training finished: Reached final iteration

Training Time

Start time: 31-Dec-2018 11:48:15

Elapsed time: 6 min 12 sec

Training Cycle

Epoch: 50 of 50

Iteration: 150 of 150

Iterations per epoch: 3

Maximum iterations: 150

Validation

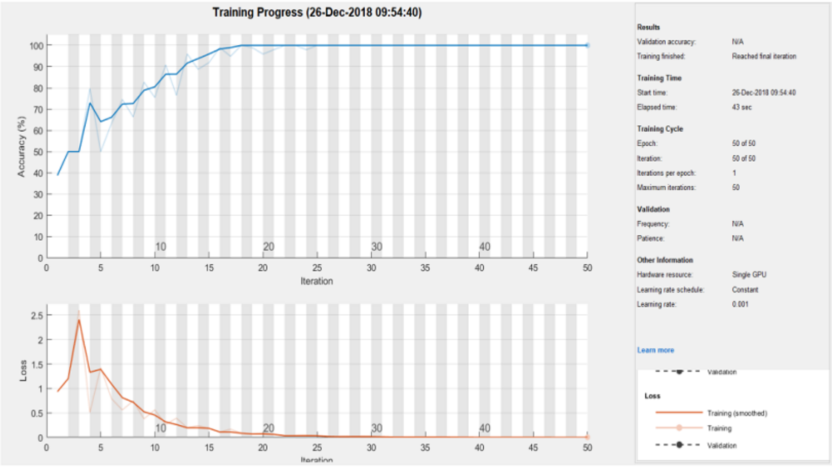
Frequency: N/A

Patience: N/A

Other Information

Hardware resource: Single CPU

Fig. 12 LSTM training progress.



Results

Validation accuracy: N/A

Training finished: Reached final iteration

Training Time

Start time: 26-Dec-2018 09:54:40

Elapsed time: 43 sec

Training Cycle

Epoch: 50 of 50

Iteration: 50 of 50

Iterations per epoch: 1

Maximum iterations: 50

Validation

Frequency: N/A

Patience: N/A

Other Information

Hardware resource: Single GPU

Learning rate schedule: Constant

Learning rate: 0.001

Fig. 13 CNN training progress.

Table 2

Method	Description	Accuracy %
LSTM-MFCC	MFCC	100
LSTM-Spectrum	Spectrogram	100
CNN-2L	2 layer	87.5
CNN-3L	3 layer	87.5
CNN-4L	4 layer	100

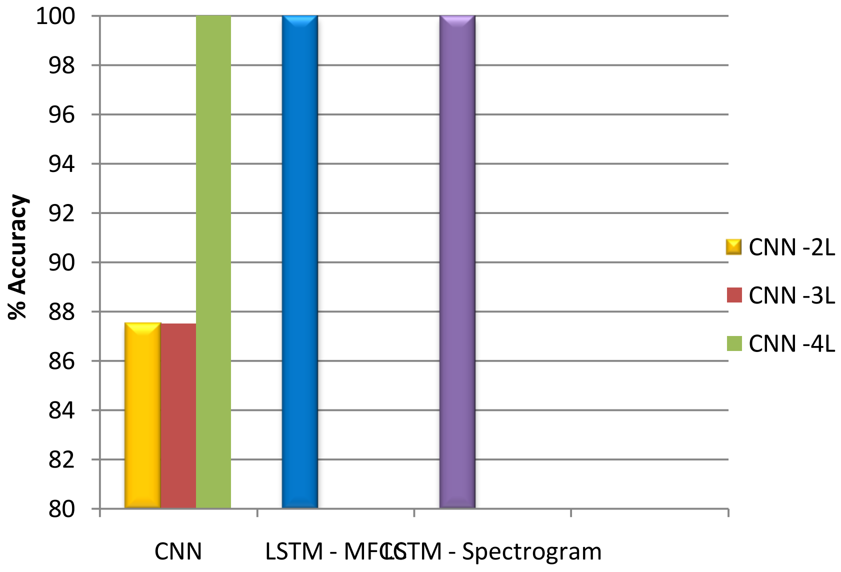


Fig. 14 Accuracy versus number of layer of CNN and the used features in LSTM

REFERENCES

[1] F. E. Abd El-Samie, "Information security for automatic speaker identification", Springer briefs in electrical and computer engineering New York: Springer, 2011.

[2] J. Eaton, N. D. Gaubitch, and P. A. Naylor, “Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost,” Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.(ICASSP), pp. 161–165, 2013.

[3] B. Cauchi, H. Javed, T. Gerkmann, S. Doclo, S. Goetze, and P. Naylor, “Perceptual and instrumental evaluation of the perceived level of reverberation,” Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP), 2016.

[4] S. Xie, N. Yan, P. Yu, M. L. Ng, L Wang, Z. Ji, "Deep Neural Networks for Voice Quality Assessment based on the GRBAS Scale", INTERSPEECH, September 8–12, 2016, San Francisco, USA.

[5] Jonathan Dennis, T Dat, and H Li, “Spectrogram image feature for sound event classification in mismatched conditions,” IEEE Signal Processing Letters, vol. 18, no. 2, PP. 130–133, 2011.

[6] Chunlei Zhang, Chengzhu yu, John H.L. Hansen,"An Investigation of Deep Learning Frameworks for Speaker Verification Anti-spoofing, JOURNAL OF LATEX CLASS FILES, August 15, 2016.

[7] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia,and A. Baskurt, "Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks", Springer-Verlag Berlin Heidelberg ICANN 2010, , pp. 154–159, 2010.

[8] y. lukic, c. vogt, o. durr, t. stadelmann, " Speaker Identification and Clustering Using Convolutional Neural Networks", IEEE international workshop on machine learning for signal processing, Sept. 13–16, 2016.

[9] S. worral, "Echo and Reverberation", based on an experiment devolved by Texas instrument Inc., 2007 and modified 2011.

[10] A. Krueger and R. Haeb-Umbach, "A model-based approach to joint compensation of noise and reverberation for speech recognition," in Robust Speech Recognition of Uncertain or Missing Data, Eds. Berlin, Heidelberg: Springer, 2011.

[11] Masashi Unoki and Sota Hiramatsu, "MTF-based method of blind estimation of reverberation time in room acoustics" 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008.

[12] M. R. Schroeder, "Natural sounding artificial reverberation", Journal of the Audio Engineering Society 10(3): 219-223, 1962.

[13] S. Nilufar, N. Ray, M. K. Islam Molla, "Spectrogram based features selection using multiple kernel learning for speech/music Discrimination". University of Alberta, Edmonton, AB, Canada.

[14] Guoshen Yu and Jean-Jacques Slotine, "Audio classification from time frequency texture," in ICASSP, pp. 1677–1680, 2009.

[15] J. Larsson, "Optimizing text-independent speaker recognition using an LSTM neural network", Master Thesis in Robotics October, 26, 2014.

[16] Xiangang Li, Xihong Wu, "Modeling Speaker Variability Using Long Short-Term Memory Networks for Speech Recognition ", INTERSPEECH , pp. 1086-1090, Sept. 6-10, 2015.

[17] V. Sze, Yu-H. Chen, T. J. Yang, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey", arXiv:1703.09039v2 [cs.CV] 13 August 2017.

[18] O. A.-Hamid, A. Elrahman Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition", IEEE/Acm Transactions on Audio, Speech, and Language Processing, VOL. 22, NO. 10, October 2014.