A Deep Autoencoder Approach To Bird Call Enhancement

Ragini Sinha, Padmanabhan Rajan
School of Computing and Electrical Engineering
Indian Institute of Technology
Mandi, India
Email: s16013@students.iitmandi.ac.in, padman@iitmandi.ac.in

***Abstract*—Due to their high performance, deep learning based approaches have attracted much attention in recent years. In this paper, we investigate deep autoencoder (DAE) based bird call enhancement. The DAE is trained utilizing layer-wise pretraining and then fine-tuned. Objective measures such as PSNR and log spectral distortion are used to compare the effectiveness of the DAE for enhancement. Furthermore, a deep neural network (DNN) based species classification system is utilized as an application to evaluate the effectiveness of the DAE. Our experiments indicate the effectiveness of the DAE based enhancement system, including the ability to provide more generalizable inputs for the DNN-based classifier. We also demonstrate the improvements obtained when band-pass filtering is performed as a preprocessing step for the DAE and the DNN.**

***Index Terms*—Autoencoder; enhancement; DNN based classification; Peak to signal noise ratio (PSNR)**

I. Introduction

In recent years, human activities have altered many aspects of the natural world [1]. To study this scientifically, systematic monitoring of landmark species in various ecosystems is required. Birds are considered a good indicator of ecosystem health. Since most birds vocalize, passive acoustic data collection can be used to monitor bird populations [2]. In general, the analysis of automatically collected acoustic data requires expertise. Moreover, the volume of data collected can be very large, rendering manual processing of the data impossible. Hence, automatic techniques to process the collected acoustic data are of much interest.

Tasks such as identification of the bird species from an audio recording are challenging due to the large inter- and intra-class variations present in bird calls. Moreover, data collected in the field are prone to noise, including sounds from the environment and other biotic sounds. For example, it is common to have the sound of the rain, the sound of a rushing river, or the sound of insects (like crickets and cicadas) when acoustic data are collected in the field. These factors make the classification task challenging [1].

Thus, it is to be expected that some form of denoising has to be performed before the species can be identified effectively. Several works on denoising of bird calls have been done in the past. If the spectral range of the noise is different from that of the bird calls, simple filtering techniques can be used. However, in many cases, there is an overlap between the frequency range of noise and bird calls, rendering simple filtering techniques ineffective. Moreover, many filtering techniques are known to induce artifacts like musical noise in the audio signal. Leveraging on the success of deep learning techniques on speech enhancement, we investigate the use of a deep autoencoder (DAE) for enhancing bird calls.

The contributions of this paper are as follows. (1) We adopt the deep autoencoder (DAE) proposed in [3] for human speech enhancement for bird calls. (2) The effectiveness of the DAE is investigated for the task of species identification, which in turn, uses a deep neural network (DNN). (3) We generate noisy versions of bird calls collected in the field using various types of noise at different SNRs. (4) We also compare the performance of the DAE with standard filtering techniques in our experimental evaluation. (5) Finally, we also perform a frequency analysis of the recordings, and then investigate simple band-pass filtering as a preprocessing step before using the DAE. Fig.1 shows the schematic block diagram for denoising of bird calls using DAE and species identification using DNN.

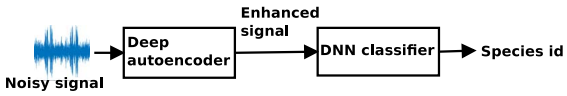

Fig. 1. Block diagram representation of species classification using DNN after denoising of bird calls using DAE.

The rest of the paper is organized as follows: section 2 discusses some of the previous ideas implemented for denoising of human speech using deep learning. Some of these can be utilized for bird calls. Section 3 discusses the general introduction to an autoencoder and section 4 discusses the proposed framework. Section 5 discusses the experimental evaluation and section 6 concludes the paper.

II. Related works

Several studies related to the enhancement of bird calls have been done earlier. For instance, Neal et al.[4] proposed to use a high-pass filter to segment bird calls in noisy environments. Bardeli et al.[5] proposed a low-pass filtering approach for detection of two bird species in a complex acoustic environment. In these studies, the choice of high-pass filter and low-pass filter depended on the range of frequencies of bird calls as well

as environmental noise. In many cases, filtering techniques can combine with other approaches to improve the noise reduction quality [6].

In recent years, deep learning based techniques have proved to be very powerful for speech enhancement. Lu et al.[3] proposed a simple autoencoder for the restoration of clean speech from noisy speech. Later Lu et al.[7] proposed a denoising autoencoder for denoising of noisy speech. Kim [8] trained three DNNs for speech enhancement and the enhanced speech is fed into a pre-trained DAE. The pre-trained DAE is used to judge the performance of each of the DNNs. Vachhani et al.[9] trained a deep autoencoder with MFCC features from clean speech for the enhancement of MFCC features of dysarthric speech. The enhanced MFCC feature is then used for a speech recognition task. Mimura et al.[10] used a joint optimization method to train a combination of DAE and DNN for noisy speech recognition. In the fine-tuning stage of the combined network, classification of phonemes is performed. Mimilakis et al.[11] proposed a recurrent neural network (RNN) to automatically learn time-frequency masks. These masks are then used to enhance spectrograms.

Narasimhan et al.[12] demonstrated simultaneous segmentation and classification of bird species using an encoder-decoder convolutional neural network and this method has shown promising results in case of in-situ recordings. In this method, the spectrograms from audio recordings are considered as images and segmentation and classification is done at the pixel level in the spectrogram image.

III. An Overview of Autoencoders for Audio Enhancement

An autoencoder is an unsupervised feed forward neural network, which learns the efficient encoding of data [13]. Typically an autoencoder is trained by back-propagation. This technique essentially allows the autoencoder weights to represent latent information as it learns to reproduce the input at its output. Fig.2 shows the schematic diagram of a basic autoencoder with one hidden layer. The following equations

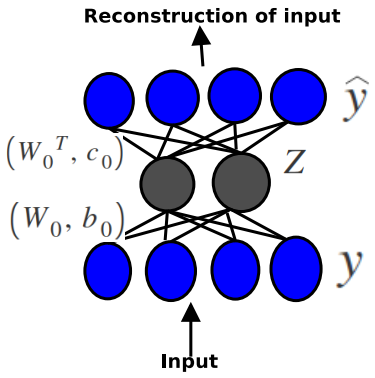

Fig. 2. Schematic diagram of a basic autoencoder. Parameters $\mathbf{W_0}$, $\mathbf{b_0}$ and $\mathbf{c_0}$ are learned at the time of training. When noisy input is given at the test time, reconstructed input is obtained at the output.

show the relation between the output of the hidden layer, $\mathbf{Z}$, the input $\mathbf{y}$ and the reconstructed input $\hat{\mathbf{y}}$ respectively [3].

$$\mathbf{Z} = \mathbf{h}(\mathbf{W_0}, \mathbf{b_0}) = \sigma(\mathbf{W_0 y} + \mathbf{b_0}) \quad (1)$$

$$\hat{\mathbf{y}} = \mathbf{g}(\mathbf{W_0}, \mathbf{b_0}, \mathbf{c_0}) = \sigma(\mathbf{W_0}^T \mathbf{Z}) + \mathbf{c_0} \quad (2)$$

Here, $\mathbf{W_0}$ represents the weight matrix of autoencoder, σ is a non-linear activation function, $\mathbf{b_0}$ is the bias vector for the input layer and $\mathbf{c_0}$ for the output layer. To learn the weight and bias parameters, the objective function shown in equation 3 must be minimized [3]. In our study, we have used an algorithm with adaptive learning rates [14] to update the weights and biases while minimizing the objective function.

$$(\mathbf{W}^*, \mathbf{b}^*, \mathbf{c}^*) \triangleq \arg\min_{W_0, b_0, c_0} \sum_{\mathbf{y}} ||\mathbf{g}(\mathbf{W_0}, \mathbf{b_0}, \mathbf{c_0}) - \mathbf{y}||_2^2 \quad (3)$$

To learn more complex latent information, more number of hidden layers are added. But, this can give rise to the problem of vanishing gradients [13]. Although, the problem of vanishing gradient can be overcome with sophisticated back-propagation method, this might require large amounts of data. An alternative to overcome the lack of large amounts of training data is to perform layer wise pre-training as adopted in [13]. In layer-wise pre-training, each layer is trained as an individual basic autoencoder. After pre-training all the layers, they are unrolled and stacked together (forming a deep autoencoder) and then fine-tuned.

IV. Proposed Framework

A. Encoder-Decoder Architecture

We adopt the architecture in [3], which is used for the enhancement of speech signals. The proposed architecture consists of 3 hidden layers. Hyper-parameters like the size of the various layers are learned using a validation set. We represent the bird call using its Mel-spectrogram. To capture temporal modulations effectively, for each frame, a context window of 7 frames is used (3 frames behind and 3 frames ahead of the current frame.) The weights of the DAE are learned by using clean Mel-spectrogram frames with context.

Fig.3 shows the architecture of our model. The size of the first and second hidden layers is set to 256 and the size of the third hidden layer is 128. We use the idea of pre-training [3]. After training each layer individually, all the layers are unrolled and stacked together for fine-tuning using back-propagation to further minimize the error. In our study RMSprop optimizer is used with a learning rate of 0.001. The batch-size is set to 512 frames and sigmoid is used as activation function for each layer. In the pre-training stage, each layer is trained till convergence and the same is followed for fine-tuning of the network.

B. Feature Extraction and Context Embedding

We extract Mel-spectrogram features from bird calls using 40 Mel filter banks. The number of FFT bins is fixed to 1024 and the frame length is set to be 20ms with a shift of half of the frame length. Context formation is obtained by considering

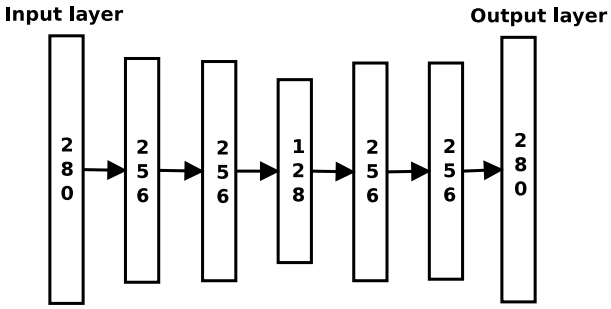

Fig. 3. Architecture of DAE model

3 frames to the left and right of the current frame. These 7 frames are concatenated to form a $40 * 7 = 280$ dimensional input feature vector, which is provided to the input layer of the autoencoder.

V. Experimental Evaluations

The experiments are conducted on data collected at the Great Himalayan National Park (GHNP) in Himachal Pradesh. This dataset is termed as GHNP dataset. This dataset consists of calls from 26 different species at 44100Hz sampling rate. The average duration of the recordings is 40 seconds per class. We utilize 14 seconds of data from each class to train the autoencoder. To create a noisy version of the data from clean data, noise is artificially added at various SNRs using FaNT [15]. Three types of noise, namely cicada, waterfall, and cricket are added at 0dB, 5dB and 10dB SNRs.

An example of the result obtained from DAE is shown in Fig.4.

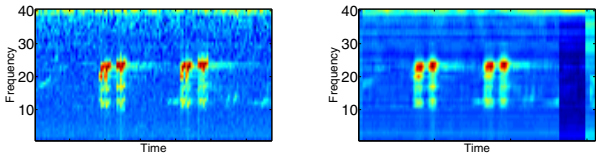

Fig. 4. The left image shows a noisy Mel-spectrogram of waterfall noise at 10dB and the right image shows the corresponding enhanced Mel-spectrogram.

The effectiveness of the DAE for call enhancement is determined through two objective measures: namely peak signal to noise ratio of Mel-spectrograms considered as images, and through the computation of log-spectral distortion.

A. Peak to Signal Noise Ratio (PSNR)

PSNR (in dB) is computed as:

$$\text{PSNR} = 10 \log_{10} \frac{\text{MAX}_i{}^2}{\text{MSE}} \tag{4}$$

Here, MSE indicates the mean squared error between the reference image and the given image and MAX_i is the maximum possible pixel value in the image. PSNR values are computed on the Mel-spectrogram images. The clean Mel-spectrogram is considered as the reference image. The PSNR values of the noisy, as well as the enhanced Mel-spectrogram (from the DAE), are computed against the reference image. This is done for a collection of about 50000 input frames. PSNR is computed for various SNRs ranging from 0 to 10dB. Box plots of the PSNR values with respect to SNR are provided in Fig5 for cicada noise, for both noisy as well as enhanced spectrogram images. It can be inferred that the DAE provides improved PSNR values at all SNRs. Similar plots are obtained for the remaining noise types.

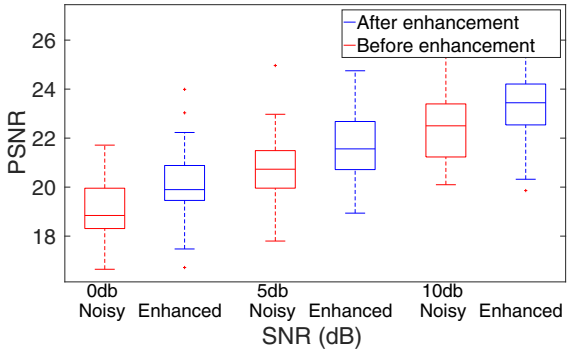

Fig. 5. PSNR plot for cicada noise for before enhancement and after enhancement using DAE.

B. Log-spectral Distortion Measurement

We also computed the log-spectral distortion between noisy and clean Mel-spectra, and between enhanced and clean Mel-spectra. Fig6 shows the histograms of the log-spectral distortion of noisy and enhanced data for cricket noise at 10 dB SNR for 11,000 frames. It can be noted from the histograms that the log-spectral distortion has a long-tailed distribution. The histogram of the enhanced spectra is shifted to the left when compared to the noisy spectra, indicating the effectiveness of the DAE.

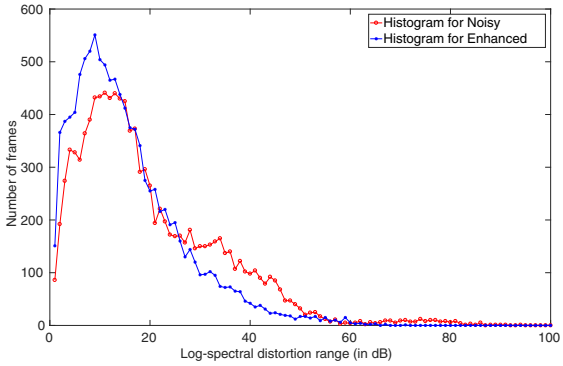

Fig. 6. Histogram for log-spectral distortion of noisy and enhanced spectra computed over 11000 frames of each.

C. Species Identification Studies

To evaluate the effectiveness of DAE based enhancement, a DNN-based species identification system for 26 species is

utilized [16]. The DNN uses 39 dimensional Mel frequency Cepstral coefficients (MFCC), delta, a derivative of delta and log-energy features. For every frame, the DNN uses the context of 7 previous and 7 next frames, resulting in a total of 15 stacked frames. The dimension of the input feature vector is $39 * 15$, i.e 585. The DNN is trained on data derived from clean audio signals. The DNN is trained using 14 seconds of data from each class using clean audio signals. While testing, MFCC features are extracted from the enhanced Mel-spectrograms obtained from the DAE. We test the performance of DAE on the three different noise types at the various SNRs given earlier.

Fig7 compares the classification accuracy of the noisy signal (without using DAE) and the enhanced signal (using DAE) for various noise types and SNRs. It can be seen that for all noise types, the classification accuracy increases with SNR. The enhancement with the DAE provides the improvement in all cases. Relative improvements in classification for cicada noise are 5.48% at 0dB, 1.82% at 5dB and 9.12% at 10dB. Similarly, for the waterfall noise, we get improvement as 4.87%, 17.93% and 9.12% respectively. For the cricket noise we have 9.42%, 11.85% and 4.25 % respectively. It can also be seen that the DAE and the DNN seem to be most effective against cricket noise. This could be a result of the reduced spectral overlap between cricket calls and bird calls.

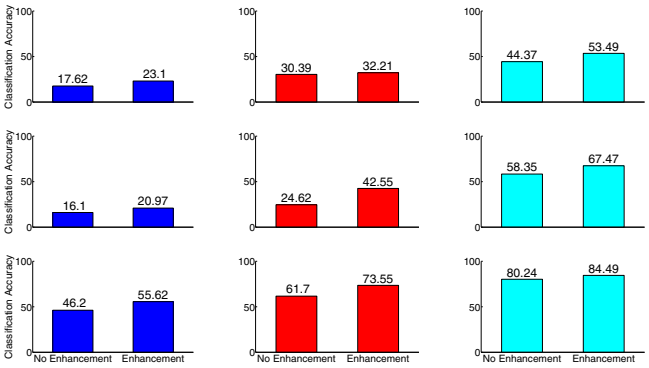

Fig. 7. Each bar plot shows the percentage classification accuracy before enhancement and after enhancement. Row 1 indicates accuracy for cicada noise, row 2 for waterfall noise and row 3 for cricket noise. For each noise type, leftmost bar graph shows accuracy at 0dB, middle one at 5dB and rightmost at 10dB.

D. Comparison with Baseline Speech Enhancement Methods

In general, log-MMSE [17] and spectral subtraction [18] are two baseline methods for speech enhancement. To investigate the effectiveness of these two methods of enhancing bird calls and classifying them, we perform the following experiment. Noisy bird calls were enhanced using these methods and the enhanced spectrograms were utilized to derive MFCC features which were fed to the DNN for classification. Fig8 shows the comparison of these methods for cricket noise. It can be seen that log-MMSE and spectral subtraction provide enhanced spectrograms which give very poor classification performance when utilized by the DNN. For log-MMSE we get 7.54% classification accuracy at 0dB, 6.99% at 5dB and 10.63% at 10dB. Similarly, for spectral subtraction we get 8.10% classification accuracy at 0dB, 6.27% at 5dB and 6.98% at 10dB. One possible reason for this poor performance could be that the spectral characteristics of the enhanced signals are altered significantly by these two methods. Another reason could be the production of artifacts by these methods, resulting in very different conditions of the spectrum on which the DNN is trained on. It is to be noted that this kind of degradation is not present in the DAE based method, which is able to provide enhanced Mel-spectrograms from which the DNN is able to utilize useful information. It can also be seen from Fig 8 that performing no enhancement (i.e. utilizing the noisy signal) provides much better classification performance than the baseline speech enhancement methods. Similar trends are observed for the other noise types.

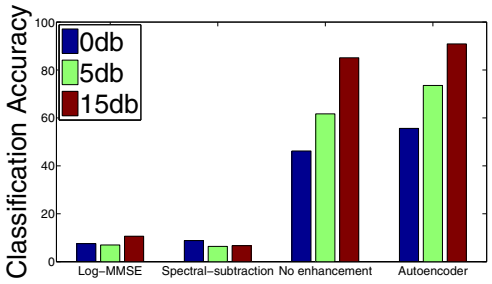

Fig. 8. Classification accuracy between baseline enhancement methods, no enhancement and DAE-based enhancement for cricket noise.

E. Band-pass Filtering as Preprocessing Before DAE

In the previous section, the DAE utilized the entire spectrum from 0 to 22.05 KHz (half of the sampling frequency.) Fig.9 shows the histogram of the frequency content of the recordings present in the GNHP dataset. From this histogram, it is clear that most of the bird calls lie within the frequency range of 0 to about 11 KHz. We wished to investigate if the performance of the DAE could be improved by restricting the enhancement to this frequency range. To achieve this, an IIR band-pass filter of order 7 is designed and applied on the recordings to attenuate the frequencies outside this desired frequency range. The lower cutoff frequency for the band-pass filter is taken as 500Hz and upper cutoff frequency is 11000Hz. After the band-pass filtering, a new DAE is trained with the reduced frequency range. Similarly, the classification system utilizing the DNN is also retrained with the MFCC features extracted from the new frequency range. Fig.10 shows the classification accuracy after performing, the band-pass filtering. It can be concluded that the band-pass filtering provides an improvement of approximately 17% at 0dB, 16.59% at 5dB and 10.19% at 10dB SNR when compared to utilizing the entire frequency spectrum. Even in the non-enhanced cases, the bandpass filtering improves classification accuracy, when compared to utilizing the entire frequency spectrum.

It is to be noted that in all cases, the DNN was trained only on the clean data. The DAE is sufficiently robust to provide

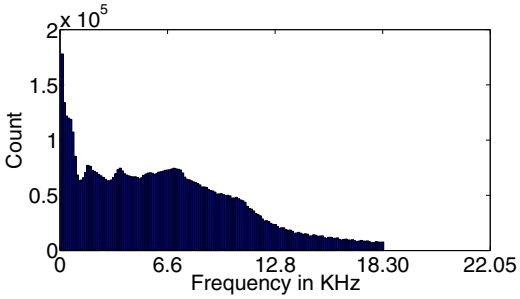

Fig. 9. Histogram of frequency content for 26 class of species from GHNP dataset.

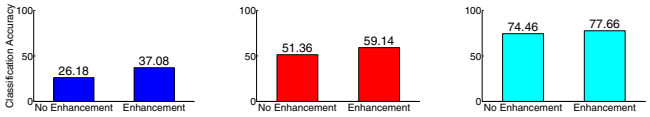

Fig. 10. Each bar plot shows the percentage accuracy before enhancement and after enhancement after band-pass filtering for waterfall noise. The leftmost bar graph shows accuracy at 0dB, middle one at 5dB and rightmost at 10dB.

enhanced features to the DNN to perform reasonably good classification.

VI. Conclusion

In this paper, we explored a deep learning based audio enhancement approach using a deep autoencoder for enhancing bird calls. Our experimental results clearly demonstrate the effectiveness of using a deep autoencoder for this task. PSNR and spectral distortion measure were utilized to quantify the amount of enhancement provided across various noise types at various SNRs. The classification accuracy of a DNN based species classification system was considerably enhanced by utilizing the DAE. When compared to traditional speech enhancement techniques, log-MMSE, and spectral subtraction, the DAE was able to provide enhanced inputs which were more closely matched to the training conditions of the classification system, which was trained only on clean data. Finally, we also demonstrated an improvement in the classification performance by restricting the spectral range using a band-pass filter, as opposed to using the entire spectrum.

References

[1] J. B. Alonso, J. Cabrera, R. Shyamnani, C. M. Travieso, F. Bolanos, A. Garca, A. Villegas, and M. Wainwright, "Automatic anuran identification using noise removal and audio activity detection," *Expert Systems with Applications* , vol. 72, pp. 83–92, 2017.

[2] T. Scott Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, vol. 18, no. S1, pp. S163–S173, 2008.

[3] X. Lu, S. Matsuda, C. Hori, H. Kashioka, "Speech Restoration Based on Deep Learning Autoencoder with Layer-Wised Pretraining," *Interspeech*, 2012.

[4] L. Neal, F. Briggs, R. Raich and X. Z. Fern,"Time-frequency segmentation of birdsong in noisy acoustic environments," *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*,pp. 2012–2015, 2011.

[5] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K. H. Tauchert, and K. H. Frommolt"Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524–1534, 2010.

[6] N. Priyadarshani, S. Marsland, I. Castro,and A. Punchihewa, "Birdsong denoising using wavelets," *PloS one*, vol. 11, no . 1, pp. e0146790, 2016.

[7] X. Lu, Y. u. Tsao, S. Matsuda and C. Hori, "Speech enhancement based on deep denoising autoencoder," *Interspeech*, pp. 436–440, 2013.

[8] M. Kim, "Collaborative deep learning for speech enhancement: A run-time model selection method using autoencoders," *arXiv preprint arXiv:1705.10385*, 2017.

[9] B. Vachhani, C. Bhat, B. Das, and S. K. Kopparapu, "Deep Autoencoder based Speech Features for Improved Dysarthric Speech Recognition," *Interspeech*, pp. 1854–1858, 2017.

[10] M. Mimura, S. Sakai, and T. Kawahara, "Joint Optimization of Denoising Autoencoder and DNN Acoustic Model Based on Multi-Target Learning for Noisy Speech Recognition," *Interspeech*, pp. 3803–3807, 2016.

[11] S. I. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller,"JA recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," *Machine Learning for Signal Processing (MLSP), 2017 IEEE 27th International Workshop on*, 2017.

[12] R. Narasimhan, X. Z. Fern, and R. Raich, " Simultaneous segmentation and classification of birdsong using CNN," *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 146–150, 2017.

[13] Li. Deng, D. Yu, and others " Deep learning: methods and applications ," *Foundations and Trends® in Signal Processing*, vol. 7, no . 3–4, pp. 197–387, 2014.

[14] T. Tieleman, and G. Hinton, " Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude ," *COURSERA: Neural networks for machine learning*, vol. 4, no . 2, pp. 26–31, 2012.

[15] H. G. Hirsch, " Fant-filtering and noise adding tool ," *Niederrhein University of Applied Sciences, http://dnt.-kr. hsnr. de/download. html*, 2005.

[16] D. Chakraborty, P. Mukker, P. Rajan and A. D. Dileep, "Bird call identification using dynamic kernel-based support vector machines and deep neural networks," *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pp. 280–285, 2016.

[17] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no . 2, pp. 443–445, 1985.

[18] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no . 2, pp. 113–120, 1979.