Image and Signal Processing of Mel-Spectrograms in Isolated Speech Recognition

Atharva Bankar
Student affiliated to MIT World Peace University

Aryan Gandhi
Student affiliated to MIT World Peace University

Dipali Baviskar
Assistant Professor affiliated to MIT World Peace University

ABSTRACT
One of the fundamental modes of communication is speech. In the past decade, many advances in the field of speech recognition system have been recorded. The conversion of acoustic waveforms into human understandable texts is the basic idea behind these systems. In this paper, an automatic speech recognition (speech-to-text) system is modelled which recognizes isolated words (one at a time). The word predictions are made based on two methods, namely Image Processing and Signal Processing. This paper presents the idea of a speech recognition system for the fundamental progress of speech recognition and also gives an overview of techniques used in each stage of speech recognition. Moreover, a comparative analysis on basis of accuracy and computation time is done. The techniques showcased in this study are used for feature extraction and then used to identify 30 spoken commands using convolutional neural networks (CNNs).

General Terms
Speech Recognition, Deep Learning

Keywords
Mel-Spectrogram, Feature Extraction, Image Processing, Signal Processing, Transfer Learning, CNNs

1. INTRODUCTION

Speech is the most common way of interaction amongst people. Due to factors like accuracy, noise filtering and computation time, the enactment of speech technologies is limited to a particular but thought-provoking range of tasks. Speech technologies developed so far have helped machines understand and convey a better understanding of human commands. With machines, vocal communication is far better than one involving keyboards and hence speech recognition systems have come under much more demand. Speech is the primary mode of communication between human beings. Therefore it is understandable that people expect voice interfaces with electronic gadgets or equipment.

Speech Recognition systems help to incorporate speech-to-text translation. An acoustic signal that holds the information of thought is the basic idea of speech. Over the last few decades, much researches have been done in the field of speech recognition systems. The most common difficulty technologists face with ASR is the challenge of mapping a sequence of acoustic measurements to the corresponding linguistic notions uniquely. This is because of vast differences in speaking rate and the various pronunciations that exist during an utterance. The processing of the audio is usually performed at three levels- Signal level, Phoneme level (basic unit of speech) and Word level processing.

This paper presents an isolated speech recognition system for a speaker-independent system and provides a comparative analysis between Signal and Image Processing based on accuracy and computation time by modeling the different isolated pronunciations spoken by different individuals using Mel-Spectrograms as a feature extraction technique.

2. LITERATURE REVIEW

The study done by T. Athanaselis and S. Bakamidis et al. 2008 [1], showcases the work on the comparison of the SVD-based noise removal scheme with the Non-Linear Spectral Subtraction (NSS) method to achieve enhanced impaired speech before feeding it to the speech recognition system. ISE is an amalgamation of the Single Value Decomposition of the input signal and a process that extracts from the signal its most active spectral properties. It was deduced that ISE is more intuitive in deciding an optimal set of parameters. At low SNR conditions, it is observed NSS performs better than ISE.

The work of V. Mitra and W. Wang et al. 2018 [2], focuses on the application of multi-view features and their discriminative transforms in deep neural network architecture for speech recognition of continuous word. Here, additional input (articulatory information) was provided to DNN which caused the error rate to be reduced by 12% relative to the baseline in both cases. The CNN-DNN architecture is efficient for combining acoustic features and articulatory features. This study proved that multi-view features along with articulatory information can improve speech recognition robustness to spontaneous and non-native speech.

The main focus of A. F. Abka and H. F. Pardede et al. 2015 [3] is a comparative study based on robustness against environmental distortions between various features extraction techniques. Log function proves to be good on clean conditions and is capable of obtaining more robust features in noisy environments. Because of the non-homomorphic property of log-based spectra, the average value of speech spectra is higher compared to log-based spectra, hence the effect of noise in low SNR can be reduced.

G. Hopper and R. Adhami et al. 1992 [4], applied Fast Fourier Transform which has been used for feature extraction for the isolated word speech recognition system. The extracted features are normalized and compared alongside previously-stored word templates. The performance of the five participants is calculated separately. Here, twenty-two utterances of each of the ten words were recorded for each individual.

A speech recognition speaker-dependent system is developed by Boussaid, L., Hassine, M. et al. 2018 [5]. Mel frequency cepstral coefficients, perceptual linear prediction and many other techniques have been used in the Feature

Extraction phase. The results include a comparative study between the hybrid approaches used.

The work of Shukla, S., Jain, M. et al. 2019 [6], is based on increasing the efficiency of speech recognition system using ANN with the help of the optimisation technique. Input in this system is dissimilar words from speakers, and the features for these inputs are extracted using Amplitude Modulation Spectrogram. The ANN is then fed with these extracted features as input for the prediction of isolated words. For optimisation Levenberg–Marquardt algorithm is used in the ANN to achieve accuracy up to 90%.

Kaur and Gurpreet et al. 2017 [7], have made use of several feature extraction techniques for the speaker-dependent speech recognition system. The results of the feature extraction techniques are analysed. The speech recognition system is built for speaker-dependent isolated words recognition.

Tabassum and Mehnaz Et al. 2017 [8], present the composition of the dynamic and instantaneous features of the speech spectrum for recognition purpose. Pitch and formant are parameters used for recognizing a word by its vowel. These two parameters are extracted and used for training the system. The system is capable of recognising and retaliating to occurrences of speech.

Tamil word classification of Lokesh and Malarvizhi Kumar et al. 2019 [9] is based on a bidirectional RNN Combined with a self-organizing map. The main aim revolves around evacuation noise from the input signal; hence the input signal is filtered using the Savitzky–Golay filter. The accuracy is further increased with the use of perceptual linear predictive coefficients. Lastly, DNN is used for the classification of the input word.

Kandagal and Amaresh et al. 2017 [10] have done a comparative study between the words and phoneme level acoustic models. The stochastic procedure is used for mounting word level and phoneme level auditory model. The results include a comparison between system performance for different size of vocabularies and the word-error rate for the above discussed two models.

A system developed by Kaur and Gurpreet et al. 2018 [11], which recognises isolated words for male and female. The feature set is developed using the MFCC feature extraction system and optimised with the use of the genetic algorithm. The extraction of features is followed by optimisation of the input; the output is fed into DNN for classification of the word along with the gender of the speaker.

The work done by Coniam, David et al. 2020 [12], is an application of speech recognition to test the oral proficiency of English Language learners. The system is trained to recognise the spoken word and analyse the output and predict the scores. A comparison is done between native and non-native speakers. The speech recognition system that has been developed is speaker-dependent.

An automatic speech recognition system is developed by M. A. M. Abu Shariah and R. N. Ainon et al. 2007 [13] to analyse, search and match input signal with the trained signals. The focus is mainly the Mel-Frequency feature extraction technique to extract useful features. The output states the spoken word in textual format.

The study of F. Itakura et al. 1975 [14], is on calculating minimum prediction residual to recognise the isolated word. The speech recognition system is speaker-dependent. A pattern to refer to a word to be recognised is stored as a time pattern of Linear Predictive Cepstrum Coefficient.

A speaker-independent automatic speech recognition system is designed by Srinivas & Nagarajan et al. 2018 [15] to predict isolated words for a moderately sized vocabulary. The feature extraction technique that has been used is 2D root cepstrum coefficients. Distance matrices are used for feature matching and KNN is implemented for classification.

The work of Paul and Dipanwita et al. 2011 [16], focuses on formant frequencies of the vocal tract and zero-crossing rate of the audio signal. The ZCR and formant frequencies used are calculated using LPC filter and signal partitioning respectively. The spoken word is then identified using neural networks.

From the literature review, it can be noticed that speech recognition systems have been built on small datasets with less than 15 words, and the number of pronunciations is not enough to build a robust automatic speech recognition system. Furthermore, no comparative study between image and signal processing for ASR is done yet. This study aims to address the aforementioned points with a robust dataset.

3. DATASET

The dataset used in this study has been acquired from Google AI compiled by the TensorFlow and Artificial Intelligence Yourself (AIY) teams. It constitutes of 30 short words viz. “Yes, No, Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Right, Left, House, Down, Up, Go, Wow, Happy, Stop, Off, On, Dog, Cat, Tree, Bird, Bed, Sheila, Marvin”. This dataset consists of 65000 one-second-long pronunciations of the aforementioned words by different individuals.

4. ALGORITHMS & TECHNIQUES

4.1 Convolutional Neural Network

This study uses the novel deep learning technique, namely convolutional neural networks (Convolutional Networks or CNN). Convolutional Neural Networks are made up of the input layer, hidden layers and output layers. The hidden layers are responsible for the decision making and competency of convolutional networks.

4.2 AlexNet

AlexNet architecture showcased by Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton et al. 2017 [17] is enclosed by eight layers; the convolutional layers make the first 5 layers of this architecture, some of these CNN layers are followed by max-pooling layers. It was observed that the saturated ReLU activation performed quite well than the standard tanh and sigmoid activations and hence has also been used in this study.

4.3 LeNet-5

LeNet-5 showcased by Y. Lecun, L. Bottou, Y. Bengio and P. Haffner et al. 1998 [18] is a CNN architecture comprising of a set of two convolutional, pooling and activation layers. This is followed by a fully-connected layer, activation, and lastly the softmax classifier.

4.4 Inception v3

The Inception v3 is a 48-layer deep CNN architecture C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna et al. 2016 [19]. It was designed to reduce the computational time to some larger extent and also reduce over fitting.

4.5 Regularization

To generalize the model on the input data, neural network architecture should be refrained from tuning its weights to a compounded model. The model can hence work better on unseen data. This study utilizes the following regularization techniques.

4.5.1 Dropout

Adding dropout in a layer randomly sets hidden units in a hidden layer to 0 at each epoch. This helps in averaging the neural network and averts the network from over fitting. Apart from improving the speed of execution of architecture, it makes the architecture to fine-tune the weights in correspondence to robust features of data.

4.5.2 Batch Normalization

This technique normalises the output of the preceding activation layer. This helps to increase the stability and speed of execution of the network. Facilitating batch normalization in the architecture permits fine-tuning the weights with a comparatively higher learning rate. Noisy data denoised by taking the difference between the batch mean and output of the preceding layer followed by dividing the result obtained by the batch standard deviation. By applying regularization, the variance of architecture is shrunk such that there is no escalation in the bias.

4.6 Convergence of Neural Networks

Convergence of convolutional neural networks on an unsuitable solution results in inconsistent outcome. When a neural network converges close to the local or global minima there is no significant improvement in the performance. To ensure that the neural network does not overshoot the minima, the following convergence techniques have been used.

4.6.1 Dynamically reducing the learning rate

The learning rate of the model is reduced by 0.5 upon no reduction in the validation loss for 5 epochs until the learning rate reaches 10^{-7}.

4.6.2 Early stopping

The training process is stopped when there is no reduction in validation loss for 20 epochs of training. This can stop the model from entering into the over fitting phase.

4.7 Mel-Spectrogram

A Mel-spectrogram symbolizes an acoustic frequency-time illustration of a sound. A spectrogram is a depiction of frequencies wavering with time and is used to denote the signal strength of audio over time. Mel-scale is a plot of pitches that is equivalently separated from each other. A Mel-spectrogram has the Mel-Scale on its y-axis.

5. PROPOSED METHODOLOGY

This study is based on pronunciations of duration equal to one-second and all the pronunciations less or greater than one-second are removed. To curb data imbalance, 1500 pronunciations for each word have been selected out of all the available pronunciations. The pronunciations have been resampled to 8000 Hz to equalize the dimensionality of the audio that would be processed and further fed to the neural networks. This paper puts front a comparative analysis of two techniques that have been used in the study, namely image processing and signal processing. The aforementioned processing techniques have been applied on Mel-Spectrograms that are obtained from audio clips of each pronunciation present in the database. Five different classes of Mel-Spectrogram of each pronunciation in the dataset has been acquired using the parameters in Table 1. The proposed methodology is displayed in Figure 1.

Table 1. Parameters of Mel-Spectrogram Class name will be used to refer to the aforementioned parameters

Class Name	Mel Bands returned	Hop Length
Class A	26	512
Class B	52	512
Class C	64	256
Class D	128	256
Class E	256	256

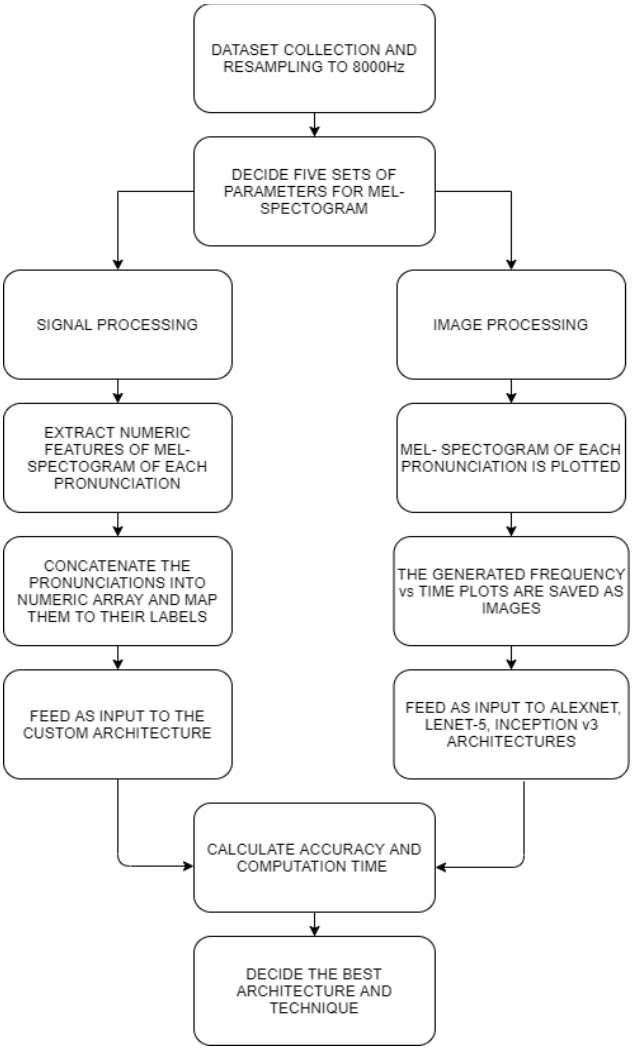

Fig 1: Flow of followed procedure

5.1 Mel-Spectrogram

In signal processing, a numeric array is computed which consists of all the pronunciations of a particular word by concatenating the Mel-Spectrogram of each pronunciation. Thus, a total of 30 arrays of 30 words are formed and each of

these 30 arrays contain 1500 pronunciations. On a similar note, all the labels of pronunciations are concatenated such that each pronunciation concatenated in the numeric array is mapped with the appropriate label. The thirty numeric and label arrays of each word are further concatenated to form one combined array of pronunciations and label respectively. These two newly formed arrays are fed to the custom convolutional network architecture as input and label. Five different custom convolutional neural network architectures have been constructed for 5 classes of Mel-Spectrograms using hyper-parameter optimization. Table 2 describes the custom ConvNet architecture used for Class-C.

Table 2. CNN architecture for Signal Processing – ClassC

Layers	**Layer Type**
L1	Convolution (64x33x1), Filters=512, Kernel =3, Activation=ReLU
	Max Pooling (2x2)
	Dropout=0.5
L2	Convolution , Filters=512, Kernel=3, Activation=ReLU
	Max Pooling (1x1)
	Dropout = 0.5
L3	Convolution Filters=256, Kernel=3, Activation=ReLU
	Max Pooling (1x1)
	Dropout=0.5
L4	Convolution Filters=256, Kernel=3, Activation=ReLU
	Max Pooling (1x1)
	Dropout = 0.5
L5	Convolution Filters=256, Kernel=3, Activation=ReLU
	Max Pooling (1x1)
	Dropout=0.5
L6	Convolution Filters=256, Kernel=3, Activation=ReLU
	Max Pooling (1x1)
	Dropout=0.5
	Convolution Filters=128, Kernel=3, Activation=ReLU
L7	Max Pooling (1x1)
	Dropout=0.5
L8	Flatten
L9	Flatten
L10	Flatten
L11	Flatten
L12	Dense (30), Activation=SoftMax

5.2 Image Processing

In image processing, a Mel-Spectrogram of each pronunciation of a word is plotted and saved as an image. Thus, each label contains images of Mel-Spectrograms equal to a number of pronunciations in it. The input size of the images that are fed in a neural network is (416, 416, 3). These images are trained on state-of-the-art architectures – AlexNet, LeNet and Inception v3 without any pre-trained weights. Figures 2 – 6 showcase the Mel-spectrogram plot of the word "Happy" with respect to each class.

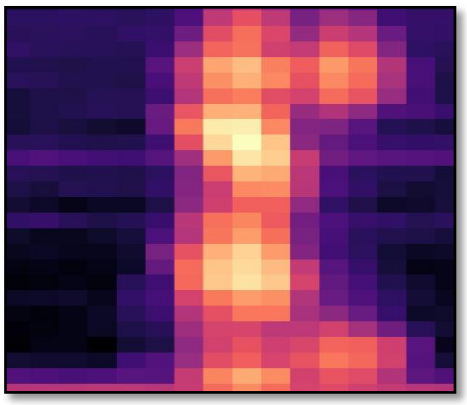

Fig 2. Class A plot for word Happy

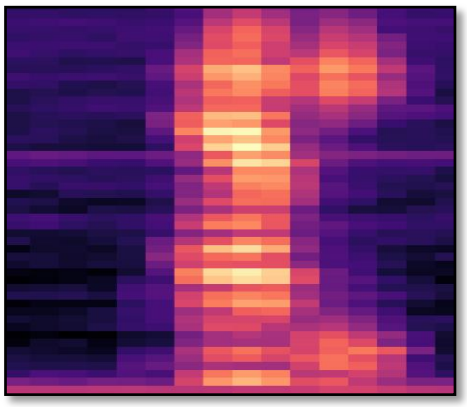

Fig 3. Class B plot for word Happy

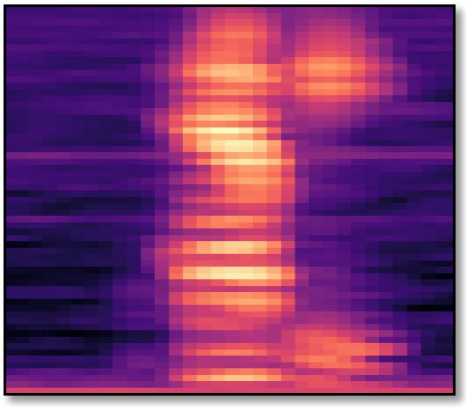

Fig 4. Class C plot for word Happy

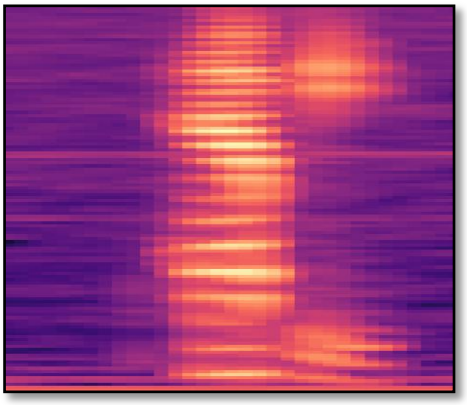

Fig 5. Class D plot for word Happy

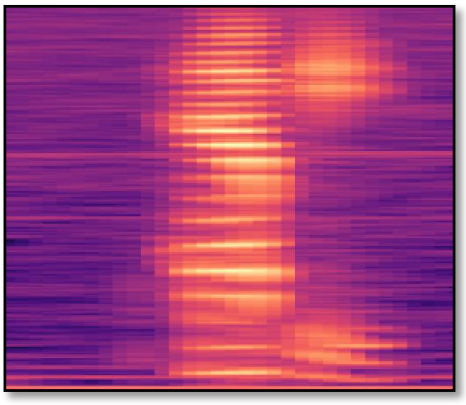

Fig 6. Class E plot for word Happy

6. RESULTS

Training of numeric arrays attained with signal processing on respective custom CNN architectures resulted in the best accuracy for class C. Table 3 showcases the accuracy and computation required for each class obtained with signal processing of Mel-Spectrograms.

Table 3. Accuracy obtained and Computation Time required for each class with signal processing of Mel Spectrograms

Class Name	Accuracy (%)	Computation Time per epoch (seconds)
Class A	90.02	8
Class B	90.50	16
Class C	91.65	38
Class D	91.09	169
Class E	91.04	220

Secondly, training of images of Mel-Spectrograms of each pronunciation of the words present in the dataset resulted in the best accuracy for Class A. Table 4 showcases the predefined architectures AlexNet, LeNet and Inception v3 which were used to deduce the accuracy and computation time for each class.

Table 4. Accuracy obtained and Computation Time required for each class with image processing of Mel Spectrograms

Class	Architecture	Accuracy	Computation Time per epoch (seconds)
Class A	AlexNet	91.35	420
	Lenet	44.26	433
	Inception v3	62.44	451
Class B	AlexNet	90.63	443
	Lenet	42.48	459
	Inception v3	55.41	467
Class C	AlexNet	91.22	496
	Lenet	41.71	505
	Inception v3	52.73	512
Class D	AlexNet	91.23	503
	Lenet	43.60	511
	Inception v3	50.51	527
Class E	AlexNet	91.03	514
	Lenet	45.60	523
	Inception v3	47.20	534

Figure 7 and 8 showcase the plots of Class A which achieved the best accuracy in image processing and Class C which achieved the best accuracy in signal processing respectively. Plots are constructed between validation accuracy and the number of epochs required for convergence of the convolutional networks. On a similar note, Figure 9 and 10 convey the validation loss in the neural network.

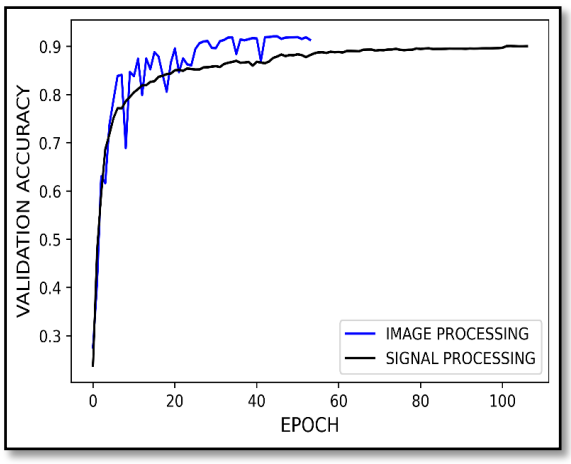

Fig 7. Validation Accuracy vs No of Epochs for Class A

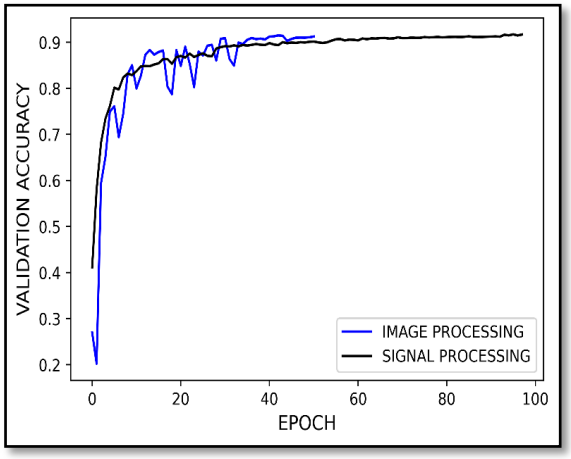

Fig 8. Validation Accuracy vs No of Epochs for Class C

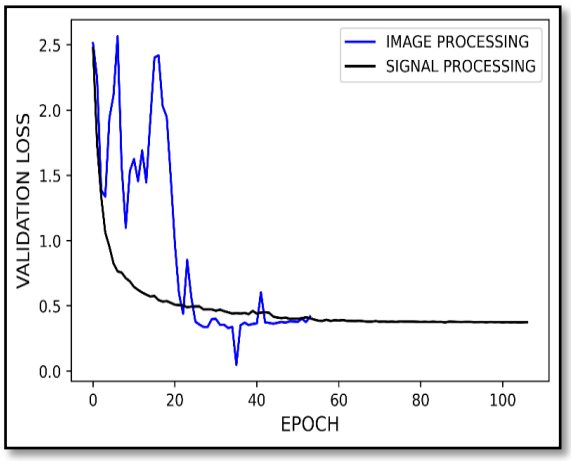

Fig 9. Validation Loss vs No of Epochs for Class A

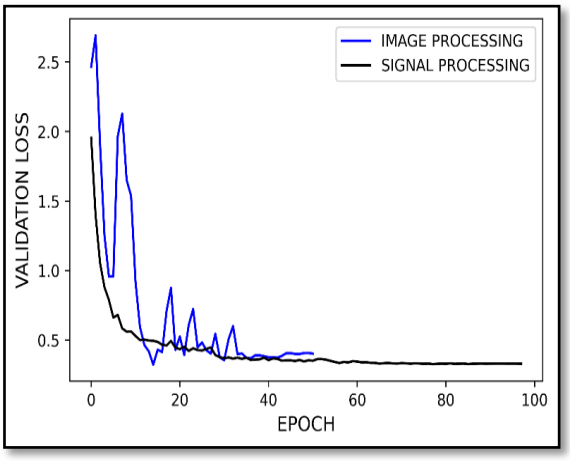

Fig 10. Validation Loss vs No of Epochs for Class C

From above-mentioned Table 3 and 4, it can be deduced that signal processing of Mel-Spectrograms gives roughly the same accuracy with much less computation time.

7. CONCLUSION

This study shows a comparative analysis between two techniques namely, signal processing and image processing for Isolated Speech Recognition on 1500 pronunciations for 30 words each. Feature extraction of pronunciations is carried out with Mel-Spectrograms to find the best trade-off between accuracy and computation time, five different classes have been created on basis of Mel Bands and Hop Length. It has been concluded that both techniques give accuracy in the range of 90% - 91% when tuned with appropriate architectures of neural network and their hyperparameters. Class C achieved the best accuracy of 91.65% with signal processing technique requiring computation time of 38 seconds per epoch, whereas class A achieved the best accuracy of 91.35% with image processing technique requiring computation time of 420 seconds per epoch.

Signal Processing is a computationally inexpensive technique as compared to Image Processing which makes it ideal for basic applications requiring Speech Recognition. Signal Processing only takes 1.90%, 3.61%, 7.66%, 33.59%, 42.80% of total-time required for Image Processing to complete one epoch for class A, class B, class C, class D and class E respectively.

8. FUTURE WORK

The researchers can further explore the combination of convolutional neural network and recurrent neural network to examine the results on the proposed system. The dataset used to train the system can be increased, hence increasing the vocabulary of the speech recognition system.

9. REFERENCES

[1] T. Athanaselis, S. Bakamidis, G. Giannopoulos, I. Dologlou and E. Fotinea, "Robust speech recognition in the presence of noise using medical data," 2008 IEEE International Workshop on Imaging Systems and Techniques, Crete, 2008, pp. 349-352, doi: 10.1109/IST.2008.4659999.

[2] V. Mitra, W. Wang, C. Bartels, H. Franco and D. Vergyri, "Articulatory Information and Multiview Features for Large Vocabulary Continuous Speech Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 5634-5638, doi: 10.1109/ICASSP.2018.8462028.

[3] A. F. Abka and H. F. Pardede, "Speech recognition features: Comparison studies on robustness against environmental distortions," 2015 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Bandung, 2015, pp. 114-119, doi: 10.1109/IC3INA.2015.7377757.

[4] G. Hopper and R. Adhami, "An fft-based speech recognition system", Journal of the Franklin Institute, vol. 329, no. 3, pp. 555-562, 1992.

[5] Boussaid, L., Hassine, M. Arabic isolated word recognition system using hybrid feature extraction techniques and neural network. Int J Speech Technol 21, 29–37 (2018). https://doi.org/10.1007/s10772-017-9480-7.

[6] Shukla, S., Jain, M. A novel system for effective speech recognition based on artificial neural network and opposition artificial bee colony algorithm. Int J Speech Technol 22, 959–969 (2019). https://doi.org/10.1007/s10772-019-09639-0.

[7] Kaur, Gurpreet & Srivastava, Mohit & Kumar, Amod. (2017). Analysis of Feature Extraction Methods for Speaker Dependent Speech Recognition. International Journal of Engineering and Technology Innovation. 7. 78-88.

[8] Tabassum, Mehnaz& Jahan, M. & Rahman, Mm & Mohamed, S. & Rashid, Mohd. (2017). Speaker Independent Speech Recognition of Isolated Words in Room Environment. International Journal on Advanced Science, Engineering and Information Technology. 7. 475. 10.18517/ijaseit.7.2.1465.

[9] Lokesh, S., Malarvizhi Kumar, P., Ramya Devi, M. et al. An Automatic Tamil Speech Recognition system by using Bidirectional Recurrent Neural Network with Self-Organizing Map. Neural Comput&Applic 31, 1521–1531 (2019). https://doi.org/10.1007/s00521-018-3466-5.

[10] Kandagal, Amaresh&Udayashankara (2017). Speaker Independent Speech Recognition Using Maximum Likelihood Approach for Isolated Words. INTERNATIONAL JOURNAL OF COMPUTER APPLICATION. 7. 10.26808/rs.ca.i7v6.10.

[11] Kaur, Gurpreet & Srivastava, Mohit & Kumar, Amod. (2018). Speaker and Speech Recognition using Deep Neural Network. International Journal of Emerging Research in Management and Technology. 6. 118. 10.23956/ijermt.v6i8.126.

[12] Coniam, David. "The Use of Speech Recognition Software as an English Language Oral Assessment Instrument: An Exploratory Study." CALICO Journal, vol. 15, no. 4, 1998, pp. 7–23. JSTOR, www.jstor.org/stable/24147601. Accessed 26 Oct. 2020.

[13] M. A. M. Abu Shariah, R. N. Ainon, R. Zainuddin and O. O. Khalifa, "Human computer interaction using isolated-words speech recognition technology," 2007 International Conference on Intelligent and Advanced Systems, Kuala Lumpur, 2007, pp. 1173-1178, doi: 10.1109/ICIAS.2007.4658569.

[14] F. Itakura, "Minimum prediction residual principle applied to speech recognition," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 23, no. 1, pp. 67-72, February 1975, doi: 10.1109/TASSP.1975.1162641.

[15] Srinivas, Nettimi& Nagarajan, Sugan & Kumar, L.s & Nath, Malaya &Kanhe, Aniruddha. (2018). Speaker-Independent Japanese Isolated Speech Word Recognition Using TDRC Features. 278-283. 10.1109/CETIC4.2018.8530947.

[16] Paul, Dipanwita& Parekh, Ranjan. (2011). Automated Speech Recognition of Isolated Words using Neural Networks. International Journal of Engineering Science and Technology. 3. 4993-5000.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (June 2017), 84–90. DOI: https://doi.org/10.1145/3065386.

[18] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.