Journal of Innovative Image Processing (JIIP) (2020)
Vol.02/ No. 04
Pages: 202-209
https://www.irojournals.com/iroiip/
DOI: https://doi.org/10.36548/jiip.2020.4.005

Analysis of Complex Non-Linear Environment Exploration in Speech Recognition by Hybrid Learning Technique

Dr. Samuel Manoharan,
Professor, Department of Electronics,
Bharathiyar College Of Engineering and Technology,
Thiruvettakudy, Karaikal, India.
Email id: jsamuel@bcetedu.in

Dr. Narain Ponraj,
Assistant Professor,
Karunya University,
Coimbatore, India.
Email id: narainpons@karunya.edu

Abstract- Recently, the application of voice-controlled interfaces plays a major role in many real-time environments such as a car, smart home and mobile phones. In signal processing, the accuracy of speech recognition remains a thought-provoking challenge. The filter designs assist speech recognition systems in terms of improving accuracy by parameter tuning. This task is some degree of form filter's narrowed specifications which lead to complex nonlinear problems in speech recognition. This research aims to provide analysis on complex nonlinear environment and exploration with recent techniques in the combination of statistical-based design and Support Vector Machine (SVM) based learning techniques. Dynamic Bayes network is a dominant technique related to speech processing characterizing stack co-occurrences. This method is derived from mathematical and statistical formalism. It is also used to predict the word sequences along with the posterior probability method with the help of phonetic word unit recognition. This research involves the complexities of signal processing that it is possible to combine sentences with various types of noises at different signal-to-noise ratios (SNR) along with the measure of comparison between the two techniques.

Keywords: *Speech Recognition, Machine Learning, Signal Process filtering techniques*

1. INTRODUCTION

Speech recognition is one of the essential applications in many signal processing applications. It is the method to analyze the verbal content speech signal to machine-understandable format as "Text". Many investigations have been carried out to translate the speech signal into the text process [1]. In different types of speech recognition, several styles of expression are getting affected. Figure 1 shows a graphical representation of the clean and noisy spectrum of the signal. Due to the lack of an audio signal, the remote word recognizers the need for the isolation of utterance. Connected words allow isolated utterances that are similar to remote words [2]. During the computer dictation, continuous speech will provide the natural speech capabilities to determine the utterance boundaries [3].

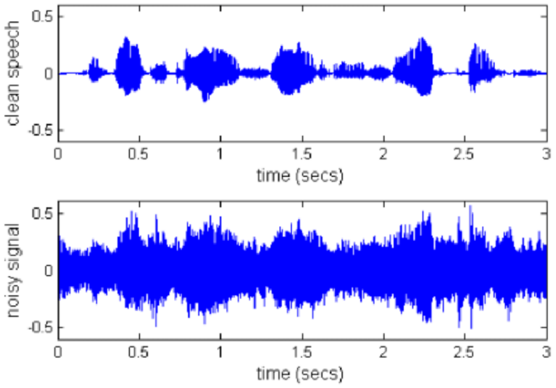


Figure 1 Graphical vision of clean & noisy signal

ISSN: 2582-4252 (online)
Submitted: 07.12.2020
Revised: 15.01.2021
Accepted: 05.02.2021
Published: 19.02.2021

And also for spontaneous speech, the speech recognition system can handle the diversity of natural speech features [4]. Figure 2 shows a simple example of speech recognition.

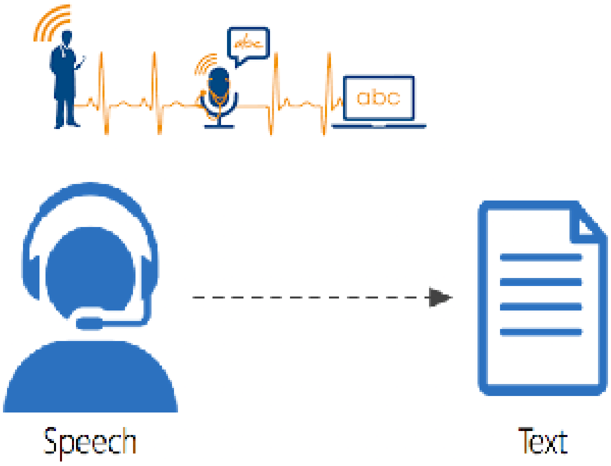


Figure 2 Example of Speech recognition

The Hidden Markov Model (HMM) is a statistical model of a dynamic Bayesian network. The states of transition conditions can be linked below. The Hidden Markov Model [5] is the output sequence of states that can be used to calculate periodic timing from a hidden point of view. In order to provide the solution to heuristic methods, speech to text decoding requires more power [6]. The rule-based HMM for speech recognition was implemented in a lot of innovative ways. The consideration of spontaneous speech and remote speech mismatches during the processing [7]. In speech recognition, there are many procedures for obtaining the best performance, such as multimodal, template-based, statistical alone, data-driven approach recognition techniques [8]. It is important to integrate the learning method for speech recognition systems into it in order to obtain knowledge-based interaction in those applications. The intelligent process can have feature extraction for further developed recognition process [9]. The researcher acknowledges in integrating phonetic word units and pattern units for the acoustic process in order to achieve the best results in the speech recognition process. Finally, with the aid of a decision-making machine, the acoustic characteristics will be measured [10]. Figure 3 shows a structure of HMM for our proposed model of statistical analysis.

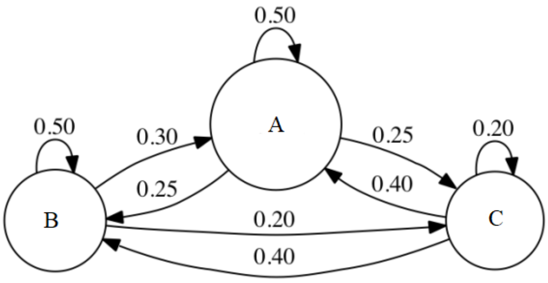


Figure 3 The structure of Hidden Markov Model

2. ORGANISATION OF THE RESEARCH

The structure of the research is organized as follows. Section 3 includes associated work with the complicated nonlinear approach of recent speech recognition. Section 4 provides the proposed structure for complex nonlinear handling. Section 5 delivers a description of the results and discussion. Finally, Section 6 concludes the research work along with the future scope.

3. RELATED WORK

The extraction of spectral resonance output by analog filter bank and logic circuits in the early years. These logic circuit analyses were investigated by Klevansand et.al [1]. Velichko et al are introducing new methods of speech recognition using discrete utterance [2]. The dynamic programming techniques for speech recognition in the early years were proposed by H.Sakoe[11].The speech or message signals are easily corrupted or disturbed by noise signals. The noise can be unpredictable and it should be avoided in the process. Many research articles discuss additive noise properties for acoustic enclosures. There is a lot of disruption even after performing the de-noising filter with many wavelet transformation methods for complex nonlinear problems in speech recognition [12].

In Figures 4 & 5, the front end and back end of the basic traditional speech recognition process are shown. The unsupervised signal processing approaches solved many problems in recognition in past decades. S. Boll et.al proposes spectral subtracts techniques in the noisy spectrum domain [13]. P. C. Loizou et.al provides the solution for the removal of noise spectrum signals with wiener filtering techniques [14].

The complex nonlinear problems are addressed by many machine learning algorithms [15]. The Convolutional Neural Network (CNN) is constructed to solve those problems in the perception task [16] and local minima and maxima activation by kernel filters approach. This method is used to extract the features effectively from the speech domain of time-frequency representation acoustic signals. The kernel function is applied over raw signals and extracted accurately [17]. Deep learning method is to reduce noise at the front end of raised problems. The construction of the acoustic model has characterized the speech distribution by Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) [6]. The common method is maximum a posterior (MAP) estimation mostly used in the noise adaptation process. All the above conventional approaches to speech recognition that are outside the reach of this survey [12] cannot be enumerated in this research. By forming a network with a neural node, the nonlinear activation function can be computed. Because this single neural node provides a lot of information with many layers in machine learning algorithms. Practically, the DNN can implement many layers that consist of many neural nodes. The input and outputs can be constructed as a very complex network for nonlinear activation function [18]. In DNN, there are many types of layers exist, such as a fully connected layer, convolution layer, and recurrent signal processing layer. The high-level representation for speech recognition can be tuned by spectrogram for higher potential computation. It is computing through many layers with an unsupervised strategy [19], [20].

3.1 PROBLEM STATEMENT

Particularly, these constructed models are very limited with the system in speech recognition. The collected data keep on changing with observed or acquired data for the speech recognition process. It also struggled to integrate large-scale data into the general model. The dynamic exploration of the nonlinear world is still a gap for further speech recognition studies.

3.2 PROBLEM SOLUTION

The observation from the previous research, there should be an improvement in simulation to realistic, solving unpredicted noise, dealing with multiple channels and vocabulary improvements to cloud storage or big data era. The need is to derive the acoustic model from the traditional to modern. The discriminative DNN is a powerful tool for the speech recognition process to achieve the highest accuracy. Recently, there is more investigation with DNN acoustic models for new noise adaptation techniques. Also, the massive collection of noisy speech data is feasible through many microphones. DNN acoustic model having a giant database of practiced factors with cloud computing provides high accuracy results [21]. These developments provide speech recognition systems to become more effective and realistic. Generally, the knowledge-based approach provides excellent results in speech recognition.

4. PROPOSED FRAMEWORK

Initially, the suggested approach is to process speech signals via an acoustic processor that can transform the vectors of the spectral features from the extracted features. It can be defined by the speech time-varying properties. The special language reference library is incorporated with many syntax, semantics, and phonetics word units which provide the best results for speech recognition. The HMM statistical probability is used to model for acoustic process and provides realization techniques to sounds and words [5]. Additionally, the language model comprises syntax, semantics and phonetics unit recognition.

4.1 AI-based approach (First Phase)

In the first phase of the proposed framework, the incoming raw speech signal will be pre-processed for the knowledge base process which is shown in figure 4. The feature extraction is used to support the good clarity for the speech signal spectrum analysis. Within the compact region of the input speech signal[22], the assumptive

ISSN: 2582-4252 (online)
Submitted: 07.12.2020
Revised: 15.01.2021
Accepted: 05.02.2021
Published: 19.02.2021

sequence of the feature vectors can be determined. This feature extraction performs three steps as follows and this feature extraction at the first phase, process after pre-processing stage. The spectrum analysis of the input signal consists of raw features on the duration of speech interval. Static and dynamic features are composed in the second stage with a detailed function vector [17]. The comprehensive feature vectors allow the recognizer module to have compact and strong features. For different and related speech signals, the optimal features can work for realistic phenomena.

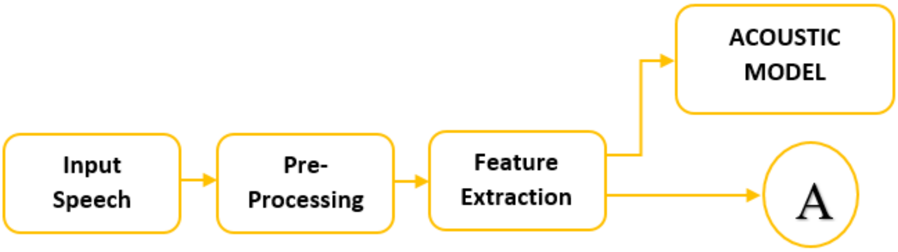


Figure 4 First phase of proposed framework

4.2 Second Phase

The language reference library comprises phonetics and pattern recognition for the word syntax and semantics to the knowledge-based acoustic model. This can understand the information based on the phonetic and linguistic style of the language [23]. Many researchers develop speech recognition systems with template-based methods which are shown effective results [24]. On the other hand, the effective development in speech sounds is based on acoustic-phonetic knowledge due to error analysis of the knowledge-based algorithm. Due to our huge amount of linguistic and phonetic word unit gives realistic pure output in many intervals. Our knowledge-based architecture is derived from a study of HMM and the artificial intelligence approach [25].

Generally, the artificial intelligence algorithm constructs the model for better work. The acoustic model with a recognition algorithm will design the concept of units of speech spectrum vectors. After processing the signal spectrum vectors, with the assistance of an acoustic model which gives recognition templates as shown in figure 5. According to the polynomial as,

$$k(\vec{x_i},\vec{x_j}) = (\vec{x_i},\vec{x_j}+1)^d$$

According to the kernel function, the transformation for the nonlinear function to linear function by a key equation as,

$$k(\vec{x_i},\vec{x_j}) = \varphi(\vec{x_i}).\varphi(\vec{x_j})$$

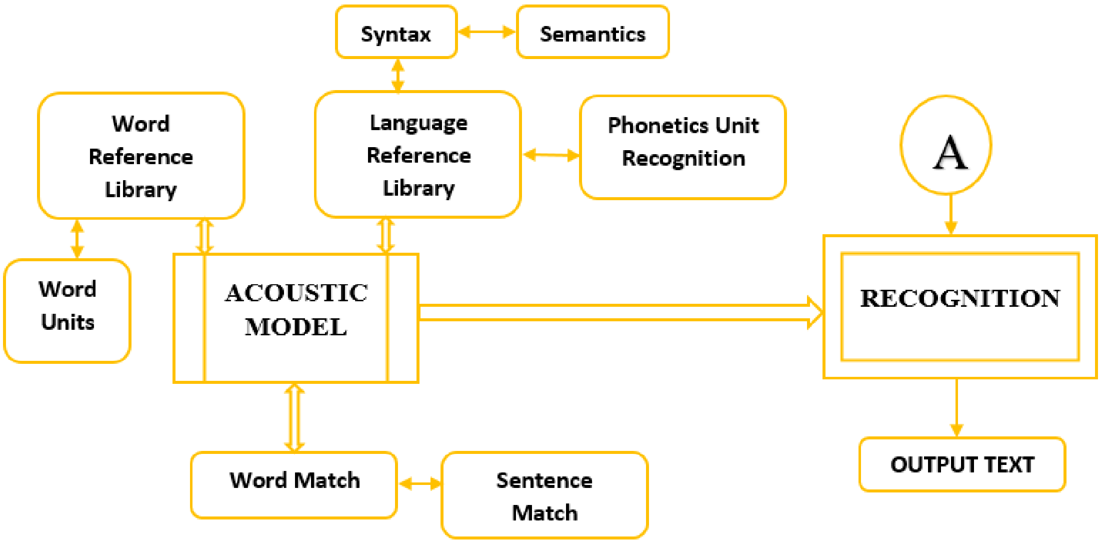


Figure 5 Acoustic model with Recognition phase

ISSN: 2582-4252 (online)
Submitted: 07.12.2020
Revised: 15.01.2021
Accepted: 05.02.2021
Published: 19.02.2021

Here, the computing sections are simply combining the statistical and learning algorithm. These distributed processes are working parallel in speech recognition processing that arranges for ultimately better results [26]. The SVM is a powerful tool in any recognition approach for spectrum analysis. By this method, the classification will be done for the recognition of the text of the input speech signals received. The linear classifier is with a huge amount of dataset functions. The classification can be computed by kernel trick as,

$$\overrightarrow{k_t}\,.\,\varphi(\overrightarrow{x_\iota}) = \sum_i \alpha_i y_j.\ k\left(\overrightarrow{x_\iota},\overrightarrow{x_j}\right)$$

The soft margin can be calculated by SVM in the form of,

$$\frac{1}{n}\sum_{i=1}^{n}\max\left(0,1-y_i(w^T.x_i-b)\right)+\delta(\|w\|^2)$$

Where, δ is the hard margin classifier for linearly classifiable input speech data. The dataset function can be varied and verified for the future fitting classifier. The co-ordinate descent algorithms solve this problem by SVM, Maximise,

$$f(c_1,\ldots c_n)=\sum_{i=1}^{n}c_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_ic_i(x_i.x_j)y_j.c_j$$

Subjected to

$$\sum_{i=1}^{n}c_iy_i=0\ \&\ 0\le c_i\le\frac{1}{2n\delta}\ for\ all\ i$$

Where $i\in\{1,\ldots n\}$
The regularization and stability calculation for the class optimization problem such as,

$$\hat{f}=\arg\underbrace{min}_{f\in H}\hat{\varepsilon}(f)+R(f)$$

Where, H is the set of hypothesis being considered.

$$R(f)=\delta_k\|f\|_H$$

If ε increases prediction and it will be developed to the perfection.

The proposed structure is used to support the changing complex non-linear function of features to control the extracted feature prediction during the training process of the dataset. Since the SVM classifier method is capable of regulating the complex non-linear function in the recognition module by kernel trick [27].

5. RESULTS AND DISCUSSION

The proposed classifier is neglects many unwanted features from the extracted output. In figure.6, the SVM method is controlling the model nonlinear complexity by using the small number of extracted output. Figure 6 (a) shows the binary vision of the estimated noisy speech input spectrum. Figure 6 (b) depicts that the removed noisy spectrum from the binary vision image. The circle in Figure 6(c) indicates that some spectrum data mismatches word content with the original real-time voice because of noisy speech signal input. Finally, the segregated output of noisy inputs is shown in figure 6 (d) respectively. Table 1 shows the comparison analysis result of speech recognition. The "grp1.wav" series files are having two various speech records in all files. Therefore 10 speech records are being tested for each classifier method with two different appending modules named word unit & phonetic unit. The HMM performs a low-efficiency rate of 76% with the word unit acoustic model. But it is obtains a high-efficiency rate of 82% with the phonetic library with the acoustic model. The proposed hybrid model also achieves efficiency rates of 88 percent and 94 percent with references to the word unit and phonetic library, respectively.

DOI: https://doi.org/10.36548/jiip.2020.4.005

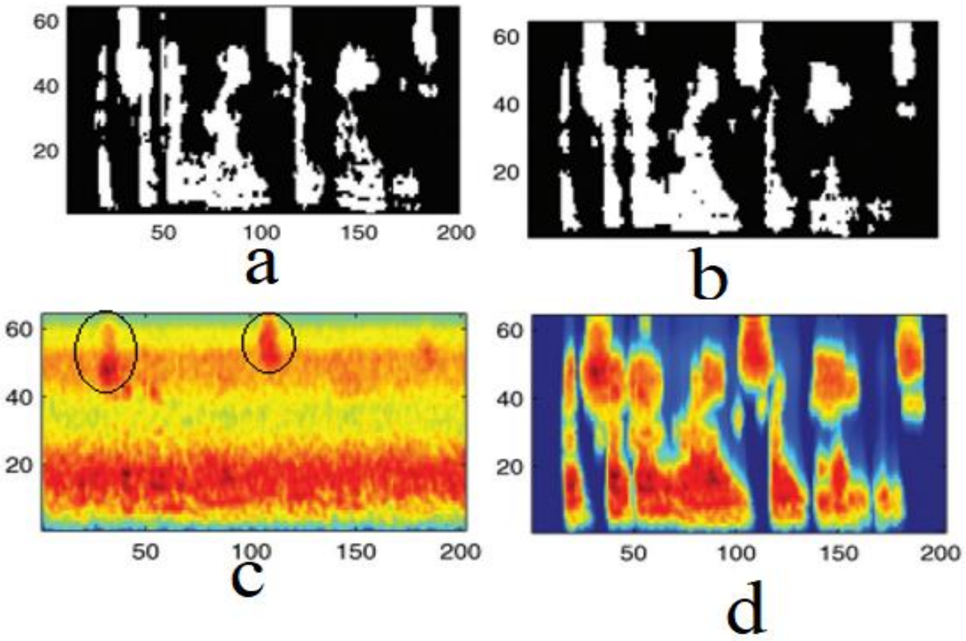


Figure 6 Obtained results from second phase of proposed framework

Generally, the speech recognition system is having to measure metrics for accuracy and speed. The Word Error Rate (WER) is one of the real-time factors that measure the accuracy of the system.

$$WER = \frac{S + D + I}{N}$$

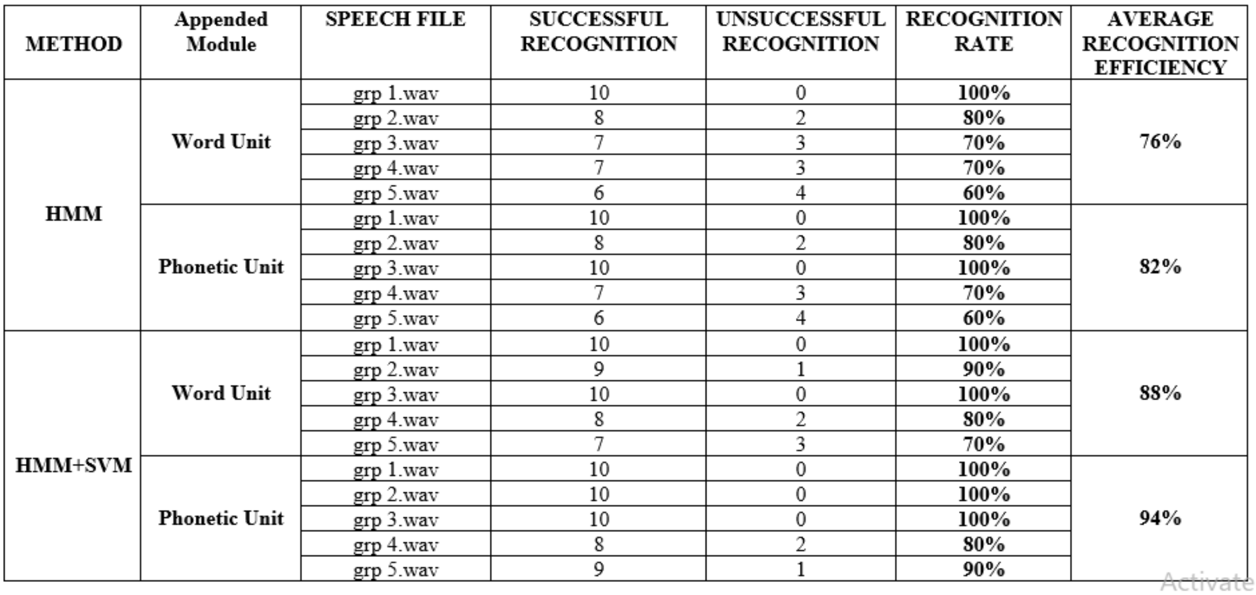
Where S is the number of substitutions, D is the number of deletions, I is the number of insertions. N is the number of total words in database for the reference. The Word Recognition Rate (WRR) is used to measure the performance of the speech recognition system.

$$WRR = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N}$$

Where H is the number of correctly recognized words. Also, the perfect accuracy detection will be determined by Single Word Error Rate (SWER) and Command Success Rate (CSR). Voice Activity Detection (VAD) is to identify the metric which it is used to detect the presence of human speech. The VAD can be evaluated by introducing noise as a speech during a calm period, front end clipping inserted noise to speech duration.

ISSN: 2582-4252 (online)
Submitted: 07.12.2020
Revised: 15.01.2021
Accepted: 05.02.2021
Published: 19.02.2021

Table 1 Comparison of recognition rate for various speech signal



METHOD	Appended Module	SPEECH FILE	SUCCESSFUL RECOGNITION	UNSUCCESSFUL RECOGNITION	RECOGNITION RATE	AVERAGE RECOGNITION EFFICIENCY
HMM	Word Unit	grp 1.wav	10	0	100%	76%
		grp 2.wav	8	2	80%	
		grp 3.wav	7	3	70%	
		grp 4.wav	7	3	70%	
		grp 5.wav	6	4	60%	
	Phonetic Unit	grp 1.wav	10	0	100%	82%
		grp 2.wav	8	2	80%	
		grp 3.wav	10	0	100%	
		grp 4.wav	7	3	70%	
		grp 5.wav	6	4	60%	
HMM+SVM	Word Unit	grp 1.wav	10	0	100%	88%
		grp 2.wav	9	1	90%	
		grp 3.wav	10	0	100%	
		grp 4.wav	8	2	80%	
		grp 5.wav	7	3	70%	
	Phonetic Unit	grp 1.wav	10	0	100%	94%
		grp 2.wav	10	0	100%	
		grp 3.wav	10	0	100%	
		grp 4.wav	8	2	80%	
		grp 5.wav	9	1	90%	

6. CONCLUSION

A single statistical model was discussed in the proposed work, including our hybrid model along with the phonetic unit. Better results are obtained by the recognition rate with the hybrid classifier process. As compared to the statistical model, the overall performance is also high in the hybrid system. The combination of self-determining and information sources is very optimal and cannot be solved easily. Many levels of phonetics, syntax, and pragmatics, and semantics, such as power spectrum disorder, disquieting execution and compilation time, can cause many problems. There is also a drawback of this proposed architecture that needs the perfect classification in a further process.

REFERENCES

[1] Sadaoki Furui, 50 years of Progress in speech and Speaker Recognition Research, ECTI Transactions on Computer and Information Technology, Vol.1. No.2 November 2005.

[2] V.M.Velichko and N.G.Zagoruyko, Automatic Recognition of 200 words , Int.J.Man-Machine Studies,2:223,June 1970.

[3] Abreu Araujo, F., Riou, M., Torrejon, J. *et al.* Role of non-linear data processing on speech recognition task in the framework of reservoir computing. *Sci Rep* 10, 328 (2020). https://doi.org/10.1038/s41598-019-56991-x

[4] Wei, Yixuan & Zhang, Xingxing & Shi, Yong & Xia, Liang & Pan, Song & Wu, Jinshun & Han, Mengjie & Zhao, Xiaoyun. (2017). A review of data-driven approaches for prediction and classification of building energy consumption. Renewable and Sustainable Energy Reviews. 82. 10.1016/j.rser.2017.09.108.

[5] Rabiner, L.. “Statistical Methods for the Recognition and Understanding of Speech 1.” (2004).

[6] Furui, Sadaoki. (1997). Recent Advances in Robust Speech Recognition.

[7] Shrawankar, Urmila & Thakare, V. M.. (2010). Noise Estimation and Noise Removal Techniques for Speech Recognition in Adverse Environment. 336-342. 10.1007/978-3-642-16327-2_40.

[8] Anusuya, M. & Katti, S.. (2010). Speech Recognition by Machine, A Review. International Journal of Computer Science and Information Security. 6.

[9] Santosh, K.Gaikwad & Bharti, W.Gawali & Yannawar, Pravin. (2010). A Review on Speech Recognition Technique. International Journal of Computer Applications. 10. 10.5120/1462-1976.

[10] Amodei, Dario & Ananthanarayanan, Sundaram & Anubhai, Rishita & Bai, Jingliang & Battenberg, Eric & Case, Carl & Casper, Jared & Catanzaro, Bryan & Cheng, Qiang & Chen, Guoliang & Chen, Jie & Chen, Jingdong & Chen, Zhijie & Chrzanowski, Mike & Coates, Adam & Diamos, Greg & Ding, Ke & Du, Niandong & Elsen, Erich & Zhu, Zhenyao. (2015). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.

[11] H.Sakoe and S.Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition ,IEEE Trans.Acoustics, Speech, Signal Proc.,ASSP-26(1):43- 49,February 1978.

[12] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745– 777, Apr. 2014.

[13] 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, Apr. 1979.

[14] P. C. Loizou, Speech enhancement: theory and practice. Abingdon, UK: Taylor Francis, 2013.

[15] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016

[16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” Neural computation, vol. 1, no. 4, pp. 541–551, 1989.

[17] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5200–5204.

[18] Meliadou, Eleni & Nakou, Alexandra. (2014). Title: How does it sound? Adding dialogue in silent movies and changing the storyline with 5-year old children.

[19] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, July 2006.

[20] B. D. Womak and J. H. L. Hansen, "Improved speech recognition via speaker stress directed classification," *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Atlanta, GA, USA, 1996, pp. 53-56 vol. 1, doi: 10.1109/ICASSP.1996.540288.

[21] Ghose, Sanchita & Prevost, John. (2020). AutoFoley: Artificial Synthesis of Synchronized Sound Tracks for Silent Videos with Deep Learning. IEEE Transactions on Multimedia. PP. 1-1. 10.1109/TMM.2020.3005033.

[22] Noda, Kuniaki & Yamaguchi, Yuki & Nakadai, Kazuhiro & Okuno, Hiroshi & Ogata, Tetsuya. (2014). Audio-visual speech recognition using deep learning. Applied Intelligence. 42. 10.1007/s10489-014-0629-7.

[23] Khdour, Thair & Muaidi, PHasan & Ahmad, Ayat & Alqrainy, Shihadeh & Alkoffash, Mahmud. (2014). Arabic Audio News Retrieval System Using Dependent Speaker Mode, Mel Frequency Cepstral Coefficient and Dynamic Time Warping Techniques. Research Journal of Applied Sciences, Engineering and Technology. 7. 5082-5097. 10.19026/rjaset.7.903.

[24] Yoshida, Takami & Nakadai, Kazuhiro & Okuno, Hiroshi. (2009). Automatic speech recognition improved by two-layered audio-visual integration for robot audition. 9th IEEE-RAS International Conference on Humanoid Robots, HUMANOIDS09. 10.1109/ICHR.2009.5379586.

[25] Zhang, Zixing & Geiger, Jürgen & Pohjalainen, Jouni & Mousa, Amr & Schuller, Björn. (2017). Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments. ACM Transactions on Intelligent Systems and Technology. 9. 10.1145/3178115.

[26] Healy, Eric & Yoho, Sarah & Wang, Yuxuan & Wang, Deliang. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners. The Journal of the Acoustical Society of America. 134. 3029-3038. 10.1121/1.4820893.

[27] Arockia Dhanraj, Joshuva & Priyadharsini, S. & Aravinth, S. & Jayaraman, P. & Krishnamurthy, Balachandar & Meganathan, D. (2020). A Review on Recent Trends and Development in Speech Recognition System. Journal of Advanced Research in Dynamical and Control Systems. 12. 521-528. 10.5373/JARDCS/V12SP1/20201099.