International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.01, April 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

A Survey – Robust Speech Recognition

Pranjal Maurya[1], DayasankarSingh[2]

[1,2]Dept. of CSE, M.M.M. University of Technology Gorakhpur, U.P., (India)

ABSTRACT

Speech recognition system affected by the noise. Presence of noise may be several reasons such as background noise, environmental noise, signal noise and other sources. These contaminations can alter the important features, properties of the voice signals and degrade the voice worth and performance. This reasons a substantial destruction to computer- human interaction system. Therefore several approaches are used to remove noise from noisy speech such as stationary, non-stationary and linear, non-linear adaptive noise cancelation, total variation de-noising etc. respectively. Thereby noise reduction is very necessary part of voice recognition, so that performance of voice recognition system can be improved. This paper gives an overview of methods which helps in noise reduction from the speech signal.

Keywords- Acoustic Modeling, Ambient Noise, Feature Extraction, Robust Speech Recognition.

I.INTRODUCTION

Voice is the fast and foremost medium of communication and human voice has a specific characteristic that differentiate one from other. Therefore speech recognition is very important not only for human beings but also an automated machine for easy and natural interaction. An Automatic Speech Recognition (ASR) is a technology that allows human beings to use their voices to speak on a computer interface in a way that, resembles normal human voice conversion. Automatic Voice Recognition is a method of affirming the talker depending on the vocalization[1][2].

The ASR technology can be applied to various applications like Biometric application, Interactive Voice Response (IVR), Call Steering, Voice Dialing, Call Routing, domestic appliance control, search, simple data entry, speech-to-text processing, automatic-information retrieval, aircraft, physically disabled person and so on. On considering these high-scale and continuously increasing value of its application speech recognition is important.

However, the general problem of ASR system lies in a variety of human voice such as speaking tone, speaking rate, age, gender, environment, accent etc. and adaption of these abrupt changes are genuinely remarkable for human beings and the second difficulty is noise, must be removed for voice recognition. Additive and convolution noise are common types of noise. During transmission of the signal, additive noise is added to the speech signal and affect it while in convolution noise speech signal influenced by convolution [16].

Therefore reduction of this noise is necessary for robust recognition and interaction can be easy and natural. For eliminating the noise it is also important to have knowledge about the common procedure of recognition. This procedure is as follows –

1.1 Voice Recording

It is a process of recording voice from many speakers with the help of microphone or other hardware at 16 KHz and put in 16 bit PCM that is programmed in mono mode.

1.2 Noise Reduction

During the recording of voice there are added some noise to the signal and reduction of this noise are important without affecting its original properties. These are the techniques used for noise reduction: (i) Filtering Methods including Spectral Subtraction Method, Wiener Filtering, Signal subspace approach (SSA) (ii) Spectral Restoration based uses Minimum Mean-Square-Error Short-Time Spectral Amplitude Estimator (MMSE-STSA), and (iii) Speech-Model-Based.

1.3 Framing and Windowing

Speech signal has time variation so done in a small window. The window is a collection of samples close a frame that takes the feature measurements and conveys a sander illustration of the speech knowledge. Every window requires some speech data, called frames. Commonly, every sequentially frame overlapped with 50 % - 70% frame and range of each frame size are 10-25 milliseconds. Further, speech data and windowing function are multiplied together. Blackman, Hamming, Bartlett, Rectangular, Gaussian and many more are several types of windows that are used.

1.4 Feature Extraction

Feature extraction of speech is a significant step to produce an effectual recognition and improving performance. After framing of the speech signal, feature extraction is done in frame-by-frame fundaments. A number of methods of feature extraction are Mel Frequency Cepstral Coefficient (MFCC), Linear Prediction Cepstral Coefficient (LPCC), Wavelet, Perceptual Linear Prediction (PLP), Temporal Patterns (TRAPS), RelAtiveSpecTrA (RASTA), Independent Component Analysis (ICA), Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA), Gamma Tone Frequency Cepstral Coefficients (GFCC) used.

1.5 Acoustic Modelling

After feature extraction acoustic modeling takes place. Acoustic modeling is a process of demonstrating a statistical representation of feature extracted from the speech signal. In acoustic modeling, a robust speech recognition feedback information is also used by the recognizer to reconstruct the feature vector. Acoustic models are going through using Hidden Morkon Model (HMM), Dynamic Time Wrapping (DTW), Artificial Neural Networks (ANN), Deep Feed Forward Networks (DNN), Dynamic Bayesian Networks (DBN), Support Vector Machine (SVM), End-To-End Automatic Speech Recognition etc. [14][15].

For a robust speech recognition, it is important to remove noise from speech signal without affecting its original meaning and characteristics.

II. RELATED WORK

In recent years, many researchers worked in the field of healthy recognition of speech. Given approach proved helpful in improving recognition performance. Selective information can be used to verify the specific speaker, used in the speech signal [1].Furthermore, Context-Dependent Pre- Trained Deep Neural Networks for Large – Vocabulary Speech Recognition has been shown to be valuable for refiningcorrectness by George E. Dahl, Dong Yu, Li Deng & A. Acero, 2012 [2]. In addition, Geoffrey Hinton, & et al. stated, for improving fine-tuning, reducing the overfitting and computational time DNN is highly likely that optimal network architecture in 2012 [3]. In phoneme level recognition, integration of Deep Recurrent Neural Networks with end-to-end training and noise weight grants reliable result in 2013 and at character level speech recognition there is no need of any expressed mapping and can be done with Recurrent Neural Network in minimum processing in 2014 [4,5].

Furthermore, Yanmin Qian and et al. in 2015 stated Multi-Task Joint- Learning scheme, where two different DNN model regressive de-noising and discriminative recognition combined into one superior framework which next incorporate in robust performance [6]. Training deep RNN with the help of Connectionist Temporal Classification (CTC) objective function convergence of training model is improved. Using visual modality features, robustness of speech is improved to noise [7].

In addition, in 2016 VikramjitMitra, Julien VanHout ant et al. stated fusion of robust features and fusion of DNN system at convolution layer proved beneficial not only for Keyword Spotting but also for Channel - and noise degraded speech [8]. Further in 2017 AbhinavThanda and Shankar M Venkatesan processed their work and came up with Multi- Task Learning approach where feature mapping of audio-visual is done and checked at various level of noise, which proved useful for improving performance [9]. The Multi-Stream Hidden Markov Model has been beneficial that transformation into a formal model i.e. audio only HMM by unifying current exponent [10].

A tabular summary of related ten papers in recent years is given in the Table 1in which author name, year of publication, dataset and methodology used and outcomes mentioned.

Table 1. Summary of previous work on Robust Speech Recognition

Sr. No.	Author's Name and Year of publication	Used Dataset	Methodology Used	Outcomes
1.	LindasalwaMuda, MumtajBegam& Elamvazuthi 2010	Speech is recorded of a male and a female for a particular word	Feature extraction(MFCC), Feature matching (DTW) methods	These algorithms are helpful improving system performance and also validate the special talker on the source of specific evidence i.e. admit in the speech signal.
2.	G. E. Dahl, Dong Yu, Li Deng & AlexAcero , 2012	Business Search Dataset	Pre-trained DNN and Context-Dependent HMM model	This approach improves recognition accuracy of 5.8% to 9.2% over CD-GMMHMMs.
3.	Geoffrey Hinton, & et. al, 2012	TIMIT dataset	Deep Neural Network (DNN)	This approach decreases the problem of over-fitting and clock time taken for Fine-tuning. And showed how replacing GMMs with HMMs obtained a substantial improvement in Automatic Speech Recognition.

4.	Alex Graves, A. Rahman,& G. Hinton, 2013	TIMIT dataset.	Deep LSTM- RNN	17.7% error is obtained when trained with end-to-end methods.
5	Alex Graves & NavdeepJaitly, 2014	Wall Street Journal (WSJ) dataset	Combination of Deep LSTM- RNN and CTC	When the absence of information, Word Error rate is 27.3%, 21.9% when Information is monogram and 8.2% when language model is trigram.
6.	Y. Qian, M.Yin, Y.You and Kai Yu, 2015	Aurora 4	Multi-Task DNN structure and fusion of audiovisual features	Word Error Rate is bellowed 10%.
7.	AbhinavThanda& Shankar M Venkatesa, 2016	GRID audiovisual corpus	RNN method for modeling and fusion of audiovisual feature	Improvement in Character Error Rate = 3.29%.
8.	VikramjitMitra, Julien VanHout ant et al. , 2016	Levantine Arabic speech dataset	DNN, CNN, and TFCNN for modeling and fusion of features, feature map at output layer and also Fusion of DNN at posterior level.	Improvement in Word Error Rate = 4.1%
9	AbhinavThanda& Shankar M Venkatesa,2017	GRID audiovisual corpus	MLT-DNN Method.	Improvement in Word Error Rate = Up to 7.23%.
10.	Ahmed HussenAbdelaziz, 2017	NTCD-TIMIT corpus	DNN for LVCSR	Compare different fusion models and conclude that Multi Stream HMM gives the best result in taken experimental setups.

III. DISCUSSION AND CONCLUSION

This paper go through the robust speech recognitions and important factor in recognition speech and also focused on different techniques used by various researchers. It has been discoursed about general procedure of speech recognition including recording of speech, noise estimation and reduction, framing and windowing, feature extractions and lastly acoustic modeling and also given algorithm's name used very often respectively. In the tabular summary results of experiments done by authors t related their work respectively, have been shown.

Robust speech recognition although a challenging task to how address it, can be good performance. In this paper, we try to give a nice review how new technologies emerged and used in robust speech recognition in recent years and also in future how technologies can be enhanced improving results. Human speech has several information about speaker such as gender, age, emotion, speaking style, personality etc. and its identification is necessary for more reliable result and in future, it is expected to focus these component for natural and easy interaction between computer human.

REFERENCES

[1] Gaikwad, S. K., Gawali, B. W., &Yannawar, P. (2010). A review of speech recognition technique. *International Journal of Computer Applications*, *10*(3), 16-24.

[2] Saini, P., & Kaur, P. (2013). Automatic speech recognition: A review. *International Journal of Engineering Trends and Technology*, *4*(2), 1-5.

[3] Muda, L., Begam, M., &Elamvazuthi, I. ''Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques' in 2010, *arXiv preprint arXiv:1003.4083*.

[4] Dahl, G. E., Yu, D., Deng, L., &Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, *20*(1), 30-42.

[5] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97.

[6] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on* (pp. 6645-6649). IEEE.

[7] Graves, A., &Jaitly, N. (2014, January). Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning* (pp. 1764-1772).

[8] Qian, Y., Yin, M., You, Y., & Yu, K. (2015, December). Multi-task joint-learning of deep neural networks for robust speech recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on* (pp. 310-316). IEEE.

IJARSE
ISSN: 2319-8354

[9] Thanda, A., &Venkatesan, S. M. (2016, December). Audio visual speech recognition using deep recurrent neural networks. In *IAPR Workshop on Multimodal Pattern Recognition of Social Signals in HumanComputer Interaction* (pp. 98-109). Springer, Cham.

[10] Mitra, V., van Hout, J., Wang, W., Bartels, C., Franco, H., Vergyri, D., &Sangwan, A. (2016). Fusion Strategies for Robust Speech Recognition and Keyword Spotting for Channel-and Noise-Degraded Speech. In *INTERSPEECH* (pp. 3683-3687).

[11] Thanda, A., &Venkatesan, S. M. (2017). Multi-task learning of deep neural networks for audio visual automatic speech recognition. *arXiv preprint arXiv:1701.02477*.

[12] Abdelaziz, A. H. (2018). Comparing Fusion Models for DNN-Based Audiovisual Continuous Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(3), 475-484.

[13] Yoshioka, T., & Gales, M. J. (2015). Environmentally robust ASR front-end for deep neural network acoustic models. *Computer Speech & Language*, *31*(1), 65-86.

[14] Shrawankar, U., &Thakare, V. (2010, October). Noise estimation and noise removal techniques for speech recognition in adverse environment. In *International Conference on Intelligent Information Processing* (pp. 336-342). Springer, Berlin, Heidelberg.

[15] Al-Haddad, S. A. R., Samad, S. A., Hussain, A., Ishak, K. A., & Noor, A. O. A. (2009). Robust speech recognition using fusion techniques and adaptive filtering. *American Journal of Applied Sciences*, *6*(2), 290.

[16] Garg, K., & Jain, G. (2016, September). A comparative study of noise reduction techniques for automatic speech recognition systems. In *Advances in Computing, Communications,and Informatics (ICACCI), 2016 International Conference on* (pp. 2098-2103). IEEE.