ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 5, May 2013

Survey of Soft Computing based Speech Recognition Techniques for Speech Enhancement in Multimedia Applications

Milind U. Nemade [1], Prof. Satish K. Shah [2]

Associate Professor, Department of Electronics and Telecommunication, K. J. Somaiya Institute of Engineering and Information Technology, Sion (E), Mumbai, India [1]

Professor, Department of Electrical Engineering, Faculty of Technology and Engineering, Vadodara, India [2]

Abstract: Speech is the fundamental, most effective, reliable and common medium to communicate in real time systems. There are so many applications of speech still to be far from reality just because of lack of efficient and reliable noise removal mechanism and techniques for preserving or improving the intelligibility for the speech signals. In this paper attempt has been stepped towards surveying the methodologies for soft computing based speech recognition techniques for speech enhancement in multimedia applications. Finally we can conclude that to improve the performance of beamforming based speech recognition system, evolutionary computational algorithm (GA) optimization can be used in multimedia applications.

Keywords: Beamformer, Dynamic Time Warping, Genetic Algorithm, Artificial Neural Network, Mel Frequency Cepstral Coefficient (MFCC), HMM based classifier.

I. Introduction

Speech is the fundamental, most effective, reliable and common medium to communicate in real time systems. In market due to advancement in technology many speech communication applications based devices is available, they are cheaper and easily available. However, undesired noises in environment cause undesired effects in real time speech processing systems. Human communications and intelligent machines are suffers from the degraded performance in which they takes decision based on what it receives as a speech.

Earlier many researchers investigated and developed various approaches for noise reduction and speech enhancements. The speech enhancement is useful for storage and transmission of speech data, also it improves speech recognition based system performance where accurate identification of words and sentences can provide automation in most of the human-machine or machine-machine based interface. Speech enhancement can boost up the performance of speech recognition systems by keeping low word error rate (WER).

There are number of speech recognition systems exists, some of them integrated into task specific applications. In practical applications a robust Mandarin Speech Recognition system using neural networks applied to multimedia interfaces performs better [5]. The speech recognition used in a multimedia speech therapy system for different problems and different ages. In addition to recording the voice and analysing the recorded spoken signal, speech recognition performs identification of speech irregularities and tracking the patient progress using time frequency analysis and neural network techniques [6].

The remaining part of the paper is organized as follows: In next section II, schematic diagram of speech recognition system is explained. Section III explains existing work related to the speech recognition techniques. Section IV explains literature survey of beamforming based speech enhancement technique. Section V explains review of soft computing techniques used for speech recognition. Section VI explains commonly used performance measuring parameters for speech recognition. Finally, paper is concluded in section VII.

II. Speech Recognition System

The Speech recognition methods were first attempted in 1952 at Bell Laboratories, Davis, Biddulph and Balashek developed an isolated digit recognition system for a single speaker [1]. There are two task of speech identification, closed set or open set. Closed set identification involved identification of speech which already exists in the database otherwise it is open set speech identification task. Isolated word speech recognition required each utterance to have silence on both sides of the word while it is difficult to recognize speech in continuous word recognition [2]. Speech recognition system is used as intelligence home in personal communication applications. It is also used in banking systems [3, 4]. Fig.1 shows the schematic diagram of speech recognition system for human being.

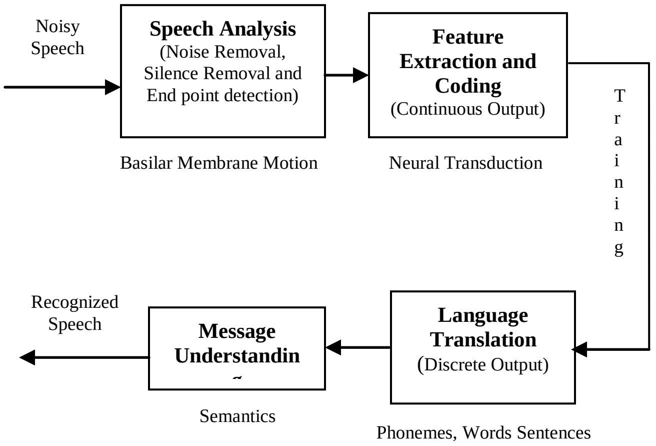

Fig. 1 Schematic diagram of speech recognition system

L. Rabiner, B.H. Juang and B. Yegnanarayana presented the approach of simple speech recognition system [7]. It consists of four main building blocks speech analysis, feature extraction, language translation and message understanding. Speech analysis stage consists of noise removal, silence removal and end point detection. End point detection and removal of noise, silence is required to improve the performance of speech recognition system. Noisy speech processes along the basilar membrane in the inner ear, which provides spectrum analysis of noisy speech. The speech analysis also deals with suitable frame size for segmenting speech signal for further analysis using segmentation, sub segmental and supra segmental analysis techniques [8].

Feature extraction and coding stage reduces the dimensionality of the input vector and maintain discriminating power of the signal. We need feature extraction because the number of training and test vector needed for the classification problem grows with the dimension of the given input. Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) are the most widely used methods for feature extraction. MFCC preferred over LPC because it is less prone to noise. The spectral signal output of speech analysis converted to activity signals on the auditory nerve using neural transduction method. Then activity signal converted into a language code within the brain, and finally message understanding is achieved.

III. Speech Recognition Techniques

Speech recognition techniques classified into three main categories as Temporal, Artificial Neural Network and Stochastic techniques. Temporal speech recognition classified as Dynamic Time Warping (DTW) and Vector Quantization (VQ), while stochastic modelling as Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) and Artificial Neural Network based speech recognition classified as Multilayer Perceptron (MLP).

DTW allows computer to find optimal match between two given speech sequences with certain restrictions. The decision to be taken is based on the global distance measures between two speech patterns [9]. For DTW there exists trade off between the recognition accuracy and the computational efficiency. In DTW optimization process is performed using dynamic programming. VQ is useful for speech coder and often applied to Automatic Speech Recognition (ASR). It uses compact codebooks for reference models. VQ used with DTW/HMM results in reduction of storage and computational time [10]. MLP is neural network technique based on back propagation (BP) algorithm used as a classifier in which nodes are connected in adjacent layer by weights. Performance of MLP degrades in the noisy environments.

Stochastic is probabilistic model deal with uncertain information and more suitable to speech recognition.HMM is popular stochastic approach characterized by finite state markov model and a set of output distributions [11]. GMM is text independent speech recognition modelling technique. In GMM each speaker has the independent GMM model and output of GMM is passed through maximum likelihood sequence detector to determine the usable speech.

IV. Speech Recognition Techniques

There are various types of advanced speech enhancement algorithms and they can be classified in main three categories, namely; filtering/estimation based noise reduction, beam forming and active noise cancellation (ANC) techniques. In beam-forming, based speech enhancement more than one speech channels (microphones) are used to process the speech. Speech signals are received simultaneously by all microphones and outputs of these sensors are then processed to estimate the clean speech signal. In adaptive beamforming, an array of antennas is exploited to achieve maximum reception in a specified direction by estimating the signal arrival from a desired direction (in the presence of noise) while signals of the same frequency from other directions are rejected. This is achieved by varying the weights of each of the sensors (antennas) used in the array. This kind of speech enhancement techniques can give better performance of the speech applications like automatic speech recognition (ASR) than signal channel processing. Only disadvantage with this class of methods is higher cost of hardware, which can put restriction on using these methods in some speech applications. The basic block diagram of beamformer is shown in Fig.2.

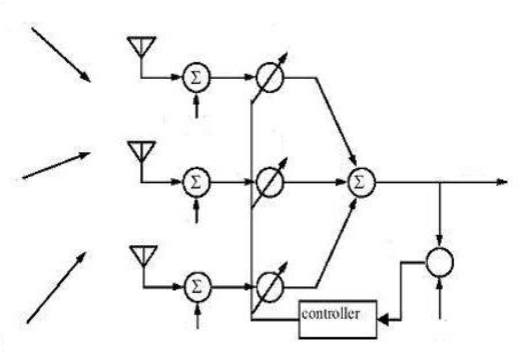

Fig. 2 Beamformer: An Adaptive array system

Frost [12] has suggested constrained minimum power adaptive beamforming, which deals with the problem of a broadband signal received by an array, where pure delay relates each pair of source and sensor. Each sensor signal is processed by a tap delay line filter after applying a proper time delay compensation to form delay-and-sum beamformer. The algorithm is capable of satisfying some desired frequency response in the look direction while minimizing the output noise power by using constrained minimization of the total output power. This minimization is realized by adjusting the taps of the filters under the desired constraint using constrained LMS-type algorithm. Griffiths and Jim [13] reconsidered Frost's algorithm and introduced the generalized side-lobe canceller (GSC) solution. The GSC algorithm is comprised of three building blocks. The first is a fixed beamformer, which satisfies the desired constraint. The second is a blocking matrix, which produces noise-only reference signals by blocking the desired signal (e.g., by subtracting pairs of time-aligned signals). The third is an unconstrained LMS-type algorithm that attempts to cancel the noise in the fixed beamformer output. For the application of hands-free speech recognition, one of the works [14] uses sequence of features to be used for speech recognition itself, to optimize a filter-and-sum beamformer instead of separating the beamformer, to be used for speech enhancement, from speech recognition system. In this work, they used frequency cepstral coefficient (MFCC) and applied to the HMM based classifier for speech recognition. Hong [15] presented the approach to improve the noisy speech recognition accuracy using noise spectral subtraction and the delay and sum beamformer based microphone array processing techniques.

V. Soft Computing Techniques

Soft computing is the collection of computational techniques in engineering disciplines which attempt to study, model and analyse very complex problems, where conventional methods fails to provide low cost solutions. The major components of soft computing are Neural Network, Fuzzy Logic and Evolutionary Computation (Genetic Algorithm).

Artificial Neural Network (ANN) is an information processing model that is inspired by the way biological nervous systems. It consist of large number of highly interconnected processing elements (neurons) working in accord to solve specific problems. ANN widely used for real time operation because ANN computations carried out in parallel. Fuzzy logic (FL) is problem solving control system methodology envisage by Lotfi Zadef. It deals with imprecise data and the data are considered as fuzzy sets. FL used in many control system applications since it mimics human control logic. Genetic Algorithms (GA's) are adaptive computational procedures modelled on the mechanics of natural genetic systems. They executed on a set of coded solutions (population) with selection/reproduction, crossover and mutation parameters [16].

In [17] authors have presented the implementation and comparison of HMM and ANN techniques for speech recognition on Field Programmable Gate Array (FPGA) chip. To obtain more precise and optimal solution authors used GA to train ANN. Results shows recognition rate for HMM is slightly better than for ANN while speed of speech recognition is much faster for ANN than HMM. Traditionally LBG algorithm is used for codebook design of vector quantization. One interesting research paper [18] presented approach based on GA-L (GA and LBG) algorithm, which improve quality of codebook for vector quantization in speech recognition systems. It is more effective than traditional LBG algorithm. One research paper [19] presented the fuzzy logic recognition approach based on power distribution pattern of a segment of a speech in real time systems. In this paper pattern generation and pattern matching process is used for real time speech processing. In application point of view delay and sum and adaptive beamforming algorithm [20] was used inside the noisy automobile environment, for the digital cellular phone application. In this paper performance criteria as signal to noise ratio and speech recognition error rate have been evaluated for the processed speech and result shows microphone array performs better than a single microphone system. In [21] it is shown that beamforming based speech enhancement technique improve the performance of speech recognition in multi-microphone environment. In this the performance of speech recognition against the filter-bank parameters; filter length and number of subbands was analysed by evaluating percentage of recognition accuracy and results obtained proved the speech enhancement capability of the beam forming technique in multi-microphone network. techniques.

VI. Performance Measuring Parameters

Speech recognition accuracy and speech recognition rate are two main terms to measure performance of speech recognition system. Speech recognition accuracy is measured in terms of Word Error Rate (WER) and speech recognition time is measured in terms of computation time. WER is a common metric of the performance of speech recognition. Here the common problem is that the recognized word sequence can have a different length from the reference word sequence.

Word error rate can then be calculated as:

$$WER = \frac{SUB + DEL + INS}{N}$$

Or

$$WER = \frac{SUB + DEL + INS}{SUB + DEL + COR}$$

Where

- SUB is the number of substitutions,
- DEL is the number of deletions,
- INS is the number of insertions,
- COR is the number of the corrects,
- N is the number of words in the reference (N=SUB+DEL+COR)

When reporting the performance of a speech recognition system, sometimes *word accuracy (WAcc)* is used instead:

$$WAcc = 1 - WER = \frac{N - SUB - DEL - INS}{N} = \frac{H - INS}{N}$$

Where

H is N-(SUB+DEL), the number of correctly recognized words.

Sometimes recognition accuracy calculated as:

$$Reco.Accuracy = \frac{Correctly\ Recognized\ word}{Total\ Recognized\ words} \times 100$$

The Real Time Factor (RTF) is a common metric of measuring the speed of an automatic speech recognition system. It can also be used in other context where an audio or video signal is processed at nearly constant rate (e.g. reading music from a CD). It is calculated as:

If it takes time *Tp* to process an input of duration *Dr* then

$$RTF = \frac{Tp}{Dr}$$

VII. Conclusion

Speech is the fundamental, most effective, reliable and common medium to communicate in real time systems. There are so many applications of speech still to be far from reality just because of lack of efficient and reliable noise removal mechanism and techniques for preserving or improving the intelligibility for the speech signals. In this paper attempt has been stepped towards surveying the methodologies for soft computing based speech recognition techniques for speech enhancement in multimedia applications. Through this review it is found that beamforming technique used widely for performance improvement of speech recognition in multimedia applications. Also we can conclude that further to improve the performance of beamforming based speech recognition system, evolutionary computational algorithm (GA) optimization can be used in multimedia applications. Essentially, we also discussed the commonly used performance measuring parameters for speech recognition.

References

[1] K. H. Davis, R. Biddulph, and S. Balashek, “Automatic Recognition of Spoken Digits,” *J. Acoust. Soc. Am.*, 24 (6): 637-642, 1952.

[2] M.G. Sumithra, M.S. Ramya, K. Thanuskodi, “Noise robust isolated word recognition” *International Conference on Communication and Computational Intelligence (INCOCCI)*, pp. 362-367, 2010.

[3] A. Burstein, A. Stolzle, and R. W. Brodersen, “Using speech recognition in a personal communications system” *IEEE International Conference on Communication*, vol.3, pp.1717-1721, 1992.

[4] T. Isobe, M. Morishima, F. Yoshitani, N. Koizumi and K. Murakami, “Voice-activated home banking system and its field trial”, *International Conference on Spoken Language*, vol.3, pp. 1688-1691, 1996.

[5] Sheu B., Ismail M., Wang M., Tsai R., “*Speech Recognition in multimedia human machine interfaces using neural networks*”, Wiley-IEEE Press, pp. 463-489, 1998.

[6] V. C. Georgopoulos, “An investigation of audio-visual speech recognition as applied to multimedia speech therapy applications”, *IEEE International Conference on multimedia computing and system*, vol.1, pp. 481-486, 1999.

[7] L. Rabiner, B.H. Juang and B. Yegnanarayana, “*Fundamentals of Speech Recognition*”, Pearson Education, first edition, ISBN 978-81-7758-560-5, 2009.

[8] H.S. Jayanna, S.R. Mahadeva, “Analysis, Feature Extraction, Modelling and Testing Techniques for Speaker recognition”, IETE Tech. Rev.,26:181-90, 2009.

[9] Bin Amin T. And Mahmood I., “Speech Recognition using Dynamic Time Warping”, *Second International Conference on Advances in Space Technologies*, pp. 74-79, 2008.

[10] S. Furui, “Vector quantization based speech recognition and speaker recognition techniques”, *Twenty-Fifth Asilmar Conference on Signals, Systems and Computers*, vol.2, pp.954-958, 1991.

[11] A. P. Varga and R.K. Moore, “Hidden Markov Model Decomposition of Speech and Noise”, *Proc. ICASSP*, pp. 845-848, 1990.

[12] O. L. Frost, III, “An algorithm for linearly constrained adaptive array processing,” *Proc. IEEE*, vol. 60, pp. 926–935, Jan. 1972.

[13] Griffiths, L.; Jim, C., "An alternative approach to linearly constrained adaptive beamforming,", *IEEE Transactions on Antennas and Propagation*, vol.30, no.1, pp. 27- 34, Jan 1982.

[14] Seltzer, M.L.; Raj, B.; Stern, R.M., "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol.12, no.5, pp. 489-498, Sept. 2004.

[15] W. T. Hong, “Residual Noise Removal on Beamforming for robust Hands-free Speech Recognition”, *International Computer Symposium (ICS)*, pp. 270-273, 2010.

[16] S.N. Shivnandam, S.N. Deepa, “*Principles of Soft Computing*”, Wiley India Pvt Ltd, Reprint: 2010.

[17] Shing T. Pan, Ching F. Chen, Jian H. Zeng, “Speech Recognition via Hidden Markov Model and Neural Network Trained by Genetic Algorithm”, *Proc. of 9th International Conference on Machine Learning and Cybernetics*, Qingdao, pp. 2950-2955, 2010.

[18] Y. Yujin, Z. Qun, Z. Peihua, “Vector quantization Codebook Design Method for Speech Recognition Based on Genetic Algorithm”, *Second International Conference on Information Engineering and Computer Science*, pp. 1-4, 2010.

[19] Tong Zhao, Peng-Yung Woo, “Fuzzy Speech Recognition”, *International Joint Conference on Neural Networks*, vol. 5, pp. 2959-2961, 1999.

[20] Stephen Oh, Vishu V., Panos P., “Hands-Free Voice Communication in an Automobile With a Microphone Array”, *IEEE International Conference on ASSP*, vol.1, pp. 281-284, 1992.

[21] Milind U. Nemade, Satish K. Shah, “Improvement in Speech Recognition Performance using Beamforming based Speech Enhancement”, *International Journal of Electronics Communication and Computer Engineering*, pp. 745-751, vol.3, Issue-4, 2012.

BIOGRAPHY

Milind U. Nemade was born in Maharashtra, India 1974. He graduated from the Amaravati University, Maharashtra, India in 1995. He received M.E (Electrical) degree with specialization in Microprocessor Applications from M.S. University of Baroda, Gujrat, India in 1999. Now he is Associate Professor and Head at Department of Electronics and Telecommunication, K.J. Somaiya Institute of Engineering and Information Technology Sion, Mumbai, University of Mumbai, India. He started PhD study at Electrical Department, Faculty of Technology and Engineering, M.S. University of Baroda, Gujrat, India. He presented and published four papers in national conferences three papers in the proceedings of international conferences and two papers in international journals. His research interest includes speech and audio processing.

Prof. Satish K. Shah is a professor in the Electrical Engineering Department at Faculty of Technology, MS University of Baroda for **last Twenty Five** years. He is a fellow of IE(I) for past Fourteen Years and has also served as the member of Committee of Vadodara local center for more than SIX years in past. He has guided more than hundred projects at UG/PG level and completed a research project on DSP based Active Power filter sponsored by AICTE, New Delhi. He has written three books on Embedded System design/ Microprocessors/ Microcontrollers and presented/published more than 25 research papers in national/international conferences/Journals. He has attended and organized several seminars, workshops, and symposiums for UGC, AICTE, IETE, and MSU.

He is a fellow of other technical associations such as: IETE, ISA and IEEE (NY) & ISTE. He has served as the member, Hon Secretary and Treasurer of their local executive committees for a span of six-eight years.