Speech Recognition: A Review

[1]Munshi Yadav, [2]M. Afshar Alam
Dept. of CSE[1, 2]
[1]Guru Tegh Bahadur Institute of Technology, New Delhi-110064
[2]School of Engineering Sciences & Technology, Jamia Hamdard, New Delhi-110062

Abstract - This paper presents fundamental concept of speech processing systems. It explores the pattern matching techniques in speech recognition system in noisy as well as in noise less environment. A study is made to review of literature for different techniques/ algorithms which are available and were developed time-to-time for the continuous improvement of technology in fast going era.

Keywords - Speech Recognition, Pattern Matching, Dynamic Time Warping Algorithm etc.

I. INTRODUCTION

Speech is the most natural means of communication of ideas and information between people. In recent years it has been realized that speech is very effective and useful medium for human-machine communication (Rahman et al., 2010). Mathematical model in communication theory has play important role in speech analysis (Shanon, C. E., 1948, 1949). First speech analysis and synthesis model named as VODER was developed in Bell Lab (Homer Dudley, 1930; Rabiner, L. R & Juang, B. H., 2006). In those days analog data was available which was very much useful for the speech recognition then after development of Pulse Code Modulation (PCM) in 1932 (W.A. Ainsworth, 1988) digitization came in the picture. In 1952 a single digit recognizer was developed for the single speaker in Bell Lab (W. A. Ainsworth, 1997).
Research in this area is growing in importance and laboratory systems have been developed which can handle a vocabulary of few hundred words. Most of these systems have been developed using a computer. Efforts have been made to develop these systems as standalone and a lot of research efforts have to be made to increase the performance of these systems.
A mathematical analysis of speech with respect to noise was done to improve the noise performance in speech communication (Schwartz and Mischa, 2008 and Rice, S. O., 1945). The proposed system recognizes the telephone quality when a single speaker speakes with a normal speed. Davis et al (1952) measured accuracy of this model in between 97% to 99%. This technique was not useful for the series of speakers.

Rest of the chapter is organized as follows. Section 2 describes about speech recognition in noisy and noise free environment. Section 3 explores about feature extraction techniques in speech recognition. Section 4 discusses about pattern matching techniques and finally chapter is concluded in section 5 followed by references.

II. SPEECH RECOGNITION

Drenthen, G. S. (2012) presented the basic speech recognition system which consists of different steps, viz., pre-processing, feature extraction, clustering and classification (Rabiner, L. R & Juang, B. H**., 2006).** In pre-processing module for input speech signal the signal-to-noise ratio is required to increase. In second step the features of the signal is extracted using feature extraction technique(s). In third step there is need to find out the centroid using k-means algorithm over feature vectors. In last step a pattern matching technique is used to recognize the speech signal. Matching score depends upon the algorithm used and size of training database (Pai, H. F., & Wang, H. C., 1993).

A computer will control the hardware operation with the help of software, a similar operation may be performed by a machine using voice command (Hansen, Per K., 1988). Zue et al. (1989) suggested that the feature extracted is the linguistic information from the input speech signal and discarding the extra information (Rahman, M. M. et al, 2010).
Zunkler, Klaus (1991) proposed an algorithm which uses the Viterbi decoding algorithm and method of hidden Markov models in combination. In this model different feature vectors assigned different weight and theoretically it was found that the different weighting schemes performed better result in speech recognition. The words that are phonetically similar create confusion in automatic speech recognition, viz., "zwei" and "drei". In basic speech recognition system confusion arises in a single phoneme, e.g., “dem" and "den"(Zunkler, Klaus., 1991). Neural Network is also useful for pattern recognition in speech (Bishop, Christopher M., 1995).
Azmi, Mohamed et al. (2008) proposed the speech recognition of Egyptian Arabic speech using syllables. The Arabic spoken words were described on the parameters as their constructing phonemes, triphones, syllables and words. Hidden Markov model toolkit (HTK) was used for the speech recognition system with the database designed by forty-four Egyptian speakers and same outperformed when used syllables. The speech recognition based on syllable performed faster than speech recognition based on word (Azmi,

Mohamed et al., 2008). Table 1 summarises the speech recognition systems/techniques/models.

Table 1: Summarises the speech recognition systems/techniques/models.

Authors	Highlights
Homer Dudley, 1930	Model presented for speech analysis and synthesis by Homer Dudley in 1930 at Bell Lab and developed a synthesizer named as VODER. In 1932 Speech recognition process improved significantly with Pulse Code Modulation (PCM).
Shanon, C. E. 1949	A mathematical model in communication theory for analysis of speech.
Davis et al, 1952	It recognizes the telephone quality when a single speaker speaks with a normal speed.
Hansen, Per K., 1988	Hardware control with the help of software and computer may also be controlled with the voice command.
Zue et al., 1989	The linguistic information from the input speech signal was considered for feature vector and discards the extra linguistic information.
Bishop, Christopher M, 1995	Neural network was used for pattern recognition in speech.
Schwartz, Mischa. 2008 & Rice, S.O., 1945.	A mathematical analysis of speech with respect to noise is done which helps to improve the noise performance in speech communication.
Azmi, Mohamed et al., 2008	A method was proposed for speech recognition of Egyptian Arabic speech using syllables.

2.1 Speech Recognition in Noisy Environment

Zhu, Q. et al. (2004) suggested a new robust adaptive speech processing algorithm and was used for estimation of speech parameters in noisy environments. In the noisy environment when noise is eliminated and system becomes noise free then with application of extended least squares (ELS) technique along with running spectrum filtering (RSF), it is possible to detect the speech characteristics correctly. The experiment in the presence of white Gaussian noise performed and results provide the robust spectrum estimation in the presence of noise (Nica, et al., 2006).

The running spectrum filters are used for designing of nonlinear running spectrum filters for accurate speech recognition system with low SNR which obstruct sometimes in efficiency in the presence of high SNR. Hayasaka, Noboru et al. (2006) suggested that use of nonlinear running spectrum filtering (NRSF) in speech processing in various cases of low-high SNRs results high accuracy and low calculation cost for the robust application (Niemann, H., et al., 2012).

2.2 Decision Fusion Technique

Al-Haddad et al. (2007) used the concept of union of the decision for speech recognition models and an algorithm for such types of speech processing, endpoint detection, MFCC, framing of the signal, vector quantization and normalization will combined to act as speech recognition. This process named as decision fusion technique used for the isolated word recognition using DTW and hidden Markov model (HMM). Experiments are performed over speech samples which are based on Malay Corpus and results observed were satisfactory 1-Nearest Neighbour (1-NN) classification method uses DTW algorithm for computing the similarity measures. When training data are large enough then 1-NN classification takes significant amount of time in computation of similarity measure. It is applicable for the limited storage, viz., embedded system where available resources are very less. Srisai, et al. (2009) proposed a novel template construction method which is based on the accurate shape averaging (ASA) technique. In this technique there is a training class available for each sequence. This technique results improvement in the performance over the 1-NN classification using DTW algorithm (Gaafar, T. S. et al., 2014).

2.3 End Point Detection In Speech Recognition

Hongbin, Gao, et al. (2009) proposed the effective endpoint detection techniques. Noise may be present due to speaker itself or due to environment or due to disturbances in the transmission system. This algorithm works on the box-counting dimension with dynamically updated threshold which improves the performance of the algorithm. The use of adaptive window will support the smooth processing of speech. Proposed algorithm requires less number of computation time and speed up the process. It is also tested over large scale of data and found the effective speech processing (Miyanaga, Y., et al., 2013).

Xu, Gang et al. (2009) proposed a new technique for endpoint detection which is based on the special feathers of Mandarin. It plays an important role in speech recognition but sometimes it is difficult to identify the position of changes of the syllables and it is not clear to find the boarder positions which are always covered by noises. The new robust algorithm used the MFCC and short - time correlation coefficient analysis which gives better results with the variation of signal to noise ratio (Kang, Guangyu Shize Guo, 2009).

2.4 Automatic Speech Recognition

Gupta, Kshitij and John D. Owens (2011) suggested that Gaussian Mixture Model (GMM) computations dominate the processing time as well as memory requirement in the automatic speech recognition systems. Graphics processors (GPU) are suitable for the exhibiting data and implementation of thread-level parallelism. Two methods were proposed for the reduction in computation. In first method using a medium-vocabulary which consists of 5k words have reduced the memory requirement by 80% without loss in accuracy with

20% overhead in computation. In second method the memory and computation savings is up to 90% and 35% respectively with 15% decrease in accuracy.

According to Kalamani, M. et al. (2014), speech processing is a unique quality to develop most promising models using which different behavior of a person in different environment can be observed.

Ali, Hazrat et al. (2014) worked on the speech recognition for Urdu language using discrete wavelets transform (DWT) and Mel frequency cepstrum coefficients (MFCC). Table 2 summarises speech recognition in noisy environment.

Table 2: Summarises Speech Recognition in Noisy Environment

Authors	Highlights
Zhu, Q. et al. 2004	A new robust adaptive speech processing algorithm developed for the estimation of speech parameters in noisy environments. After elimination of the noise use of extended least squares (ELS) technique with running spectrum filtering (RSF), it is possible to detect the speech characteristics correctly.
Hayasaka, Noboru et al., 2006	The running spectrum filters are used for designing of nonlinear running spectrum filters which results accurate speech recognition system and produces better accuracy in results and low computing cost for the robust application.
Al-Haddad et al., 2007	This algorithm uses the concept of union of the decision in the recognition models. For such types of speech processing the end-point detection, MFCC, framing of the signal, vector quantization and normalization are combined to act as speech recognition.
Srisai, Dararat, and Chotirat Ann Ratanamahatana, 2009	The proposed novel template construction technique is based on the accurate shape averaging (ASA).
Hongbin, Gao, et al., 2009	The effective endpoint detection techniques which works on the box-counting dimension with dynamically updated threshold which improve the performance of the algorithm.
Xu, Gang et al., 2009	The new robust algorithm using the MFCC and short-time correlation coefficient analysis gives better results in speech recognition with the variation of signal to noise ratio (SNR).
Gupta, Kshitij, and John D. Owens., 2011	Gaussian mixture model (GMM) computations dominate the processing time as well as memory requirement in the automatic speech recognition systems.
Ali, Hazrat et al., 2014	Speech recognition for Urdu language using discrete wavelets transform (DWT) and Mel frequency cepstrum coefficient (MFCC).
Kalamani, M. et al., 2014	Speech processing is a unique feature to develop the most promising models by which different behaviour of a person in different environment can be observed.

III. FEATURE EXTRACTION

The input speech signal passed through a feature extraction technique for extracting sequence of features which will be useful in speech recognition. There are a number of commonly used feature extraction techniques from voice samples, which are being used for speech recognition. Some of these are - analog-to-digital conversion, short-term frequency analysis, discrete Fourier transform, filter bank analysis, autocorrelation analysis, cepstral processing, linear prediction analysis and wavelet transform, etc.

3.1 Filter Bank

B.A. Dautrich et al. (1983) and L. R. Rabiner, T.B. Martin (1983) suggested that filters in filter-bank are combination of uniform and non-uniform distribution of frequency. Study was done for the combination of both as 8 numbers of uniform filters and 5 numbers of non-uniform filters in filter bank. Results have been compared with 15 uniform and 13 non uniform filter banks. Results also compare with LPC based feature and found that filter bank analysis is only 4% worse than LPC based feature and recognizer. It is proposed by B.A. Dautrich et al. (1983) and L. R. Rabiner, T.B. Martin (1983) that the performance by male speaker is better than by female speaker system. Result is better in the noiseless environment rather than to noisy environment. The performance of LPC based recognizer and by filter bank based recognizers are equal for the small vocabulary of English words/digits (Mobin, A., et al., 1989).

3.2 Linear Prediction Coefficients (LPC)

Jiang, Hai and Er. Meng Joo (2003) proposed a new method for feature extraction which includes the combination of the static LPC and h dynamic LPC and this combination will be used for feature extraction from the frame of speech. An efficient and reduced-dimension of speech has been derived for speech vector in the speech recognition system using linear discriminant analysis (LDA). The experiments are performed with hidden Markov model (HMM) as the speech recognition technique for the isolated-word recognition and result was quite efficient (Lahouti, et al., 2006).

Schroeder, Manfred R. (1982) introduced the fundamental concept of linear prediction (LP) and maximum entropy (ME) spectral analysis as well as concept of minimum cost entropy (MCE) spectral analysis and the reason of their use. In speech signal analysis role of MCE is to reduce the number of predictor coefficients which depends on arrival rates of frames. The spectral information is provided by the physical components in speech production model like lip radiation characteristics, microphone and transmission frequency response, etc. appearing at the time of steady state and slowly varying position of a particular speech utterance.

Bagul, S. G., and R. K. Shastri (2013) proposed the concept of speaker recognition which identifies a person on the basis of his/her voice. The basic idea behind this concept is to recognize and classify the voice of a particular person from the group. These features may be pitch, amplitude and frequency, etc. Using a unique identity for individual person who will obtain by features extracted from the particular speech signals and with a statistical model like Gaussian mixture model (GMM) a particular speaker can be identified. Further, it has been suggested that the speaker recognition efficiency can be improved using the fractional Fourier transform as a feature extraction technique in speaker recognition.

De Poli, Giovanni and Luca Mion (2006) proposed alternative representations of the input speech signal which is capable enough to give better performance of the speech recognition system. In the feature extraction the wavelet based features with different wavelets are used and the experiments are performed over Hindi digits recognition. The comparative study of the recognition score is done in both the conditions, with linear prediction coefficients (LPC) as well as with perceptual linear prediction (PLP) features. Using HMM as pattern matching technique is performed with speaker independent Hindi digits recognition. It is observed that the performance of recognition is 11.3% better with PLP based features than LPC based features (De Poli, Giovanni, and Luca Mion, 2006).

Xie, Lei and Zhi - Qiang Liu (2006) proposed that an audio-to-visual conversion is the main tasks in designing of speech-driven facial animation. In audio-to-visual conversion a good prediction is required to predict the facial control parameters from the acoustic speech which depends over the representation of audio features (Jiang, H, et al., 2003).
A comparative study is made using prosodic, perceptual and articulatory features in audio-to-visual conversion problem over a common task. It is observed that Mel frequency cepstrum coefficients (MFCC) give the better performance, in comparison to perceptual linear prediction coefficients (PLPC), the linear predictive cepstrum coefficients (LPCC) and the prosodic feature set.

Further studies show that the perceptual linear prediction coefficients (PLPC), the linear predictive cepstrum coefficients (LPCCs) and the prosodic feature set in combination can produce the better result in audio-to-facial animation. This different audio feature carries complementary information which is relevant to facial animation (Xie, Lei and Zhi - Qiang Liu, 2006).
Table 3: Summarises study of LPC tchniques/algorithms.

Table 3**:** Summarises study of LPC

Authors	**Highlights**
Hai, Jiang, and Er MengJoo., 2003	The improved linear predictive coding (LPC) coefficients are employed for the feature extraction where static LPC and dynamic LPC were employed as basic feature.
Schroeder, Manfred R., 1982.	The fundamental concept of Linear Prediction (LP) and Maximum Entropy (ME) spectral analysis as well as concept of Minimum Cost Entropy (MCE) spectral analysis.
Bagul, S. G., and R. K. Shastri., 2013.	Speaker recognition system using a unique identity for individual person who will obtain by features extracted from the particular speech signals and with a statistical model like Gaussian mixture model (GMM) a particular speaker can be identified.
De Poli, Giovanni, and Luca Mion., 2006.	Alternative representations of the input speech where wavelet based features with different wavelets have been used and the experiment performed over Hindi digits recognition. It has been observed that the performance of recognition is 11.3% better with PLP based features than LPC based features.
Xie, Lei, and Zhi-Qiang Liu., 2006.	Perceptual linear prediction coefficients (PLPC), the linear predictive cepstrum coefficients (LPCC) and the prosodic feature set in combination produces the better result in audio-to-facial animation since different audio features carry complementary information which is relevant to facial animation.

3.3 Optimized Feature Extraction

Lee, Chulhee et al. (1998) proposed an optimal feature extraction technique for normally distributed data in speech recognition. Optimality of the feature extraction technique suggests choosing the set of features which provide the minimum classification error. This method firstly chooses any random feature vector and finds the classification error. On the basis of observation it starts to move the feature vector in the direction where classification error may be reduced. Two search techniques are provided- namely sequential search and global search. In the first method, if more features are required then there is need to find out more feature which provides better classification accuracy in comparison to already existing features. In the second method, it will not bind to use the existing features. Performance of the proposed technique is efficient as verified by the experimental results as compare to the conventional feature extraction algorithms (Lee, Chulhee et al, 1998; Behroozmand, R., & Almasganj, F., 2005).
The feature extraction is essential in the study of pattern classification (Lee, Chulhee and Euisun Choi, 2000). The conventional feature extraction techniques are well defined between two classes or between two global function but it is not optimal for the application where multiclass problems present. A solution for optimizing the feature extraction in multiclass problems is proposed. In multiclass problems firstly explore the possibility of finding the much better feature sets which are not observed by the existing method for the

classification accuracy. Then an algorithm is required to find out such features. The performance of the proposed technique is efficient as verified by the experimental results in comparison to conventional feature extraction algorithms (Choi, E., & Lee, C., 2001).

Based on the Bhattacharyya distance the error estimation equation will support to estimate the classification error (Nehe, N. S., & Holambe, R. S., 2008). The theoretical and experimental results proved that this feature extraction technique is more powerful than other techniques rendering by Choi, Euisun, and Chulhee Lee. (2000).

Choi, Euisun and Chulhee Lee. (2003) presented a method of feature extraction using an error estimation equation which is based on the Bhattacharyya distance. The classification errors have been used as transformed feature space and criteria of feature extraction has been taken as error estimation equation. Since it has the ability to predict the error so, it is very much possible to identify the requirement of minimum number of features in the feature extraction. Experimental results show that this feature extraction technique is more helpful in comparison to other existing feature extraction methods (Cox, R. V, et al., 2000).

Midorikawa et al. (2010) proposed a solution which improves the speech recognition system in noisy environment. Since the performance of speech recognition is poor in the noise and it works better in noiseless environment. Fourier analysis is used by number of researchers but it works for the frequency component only. Specific noise frequency is reduced using general noise filters. To remove the noise effect, it is suggested to apply the wavelet transform in speech recognition. Using the cepstral analysis in noisy environment with wavelet transform and weighting coefficients improve speech recognition system.

Fleissner et al., (2011) presented acoustic classification of adaptation of non-native spoken language. The non-native speech concept is adapted in speech dialogue system by non-native English speakers. Table 4 summarises study for feature extraction techniques.

Table 4: Summarises study of feature extraction

Authors	Highlights
B.A. Dautrich et al. &L. R. Rabiner, T.B. Martin, 1983	Study is done for the number of filters in filter- bank as combination of uniform and non-uniform distribution of frequency. It is also observed that the performance by male speaker is better than by female speaker system.
Lee, Chulhee et al., 1998	An optimal feature extraction technique for normally distributed data in speech recognition.
Lee, Chulhee, and Euisun Choi, 2000	Solution for optimizing the feature extraction in multiclass problems.
Choi, Euisun, and Chulhee Lee, 2000	Feature extraction technique for multimodal data which is based on the Bhattacharyya distance.
Choi, Euisun, and Chulhee Lee, 2003	Method of feature extraction using an error estimation equation which is based on the Bhattacharyya distance.
Midorikawa et al., 2010	A solution which improves the speech recognition system in noisy environment. Specific noise frequency is reduced using general noise filters. The wavelet transform remove the noise effect in speech recognition.
Fleissner et al., 2011	The non-native speech concept is adapted in speech dialogue system by non-native English speakers.

IV. PATTERN MATCHING

The feature extracted in feature extraction using any feature extraction technique like filter bank analyser, linear prediction coefficient or Mel frequency cepstrum coefficient are stored as sequence of measurement is called a pattern. Numbers of patterns corresponding to each spoken word are used to make a database used as reference pattern(s)/template(s) (Bhattacharjee, U & Sarmah, K., 2013). The unknown test pattern is compared with reference pattern and a measure of similarities which is generally distance between the test pattern and reference pattern is compared. On the basis of the calculated distance between the test and reference pattern it decides that which reference pattern is best match with the unknown test pattern (Duda, R. O., 2012; Yadav, Munshi and Afshar Alam., 2016).

4.1 The Euclidean Distance

If $X = (x_1, x_2, x_3 \ldots x_n)$ and $Y = (y_1, y_2, y_3, \ldots y_n)$ are the feature vectors then the Euclidean distance, d between both the vectors is defined as, $d = \text{sqrt}\,[(x_1-y_1)^2 + (x_2-y_2)^2 + \ldots + (x_n-y_n)^2]$

4.2 City Block Distance

The city block distance between two vectors of size n is computed by getting summation of (x_i-y_i) where, i=1 to n. It has the value zero or more. For ideal matching distance being zero and value more than zero is use to represent poor matching/recognition.

4.3 Code Book Recognition

It is a technique for isolated word/pattern recognition developed by Burton, D. et al, (1983). In this technique instead of time normalization, it dispenses with both time and information. This is based on the vector quantization. A codebook is design for each word from the number of repetition of words in the vocabulary. When an unknown word is received vector quantization is performed with each code book and the unknown word is classified by choosing that codebook which gives the lowest average distortion per frame of speech.

4.4 Vector Quantization

It is older technique which was originally used for data compression. The density matching property is used in this technique. It is applicable for identifying the density of large and high-dimensioned data. It divides a large set of points/vectors into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms. Since data points are represented by the index of their closest centroid, commonly occurring data have low error and rare data with high error. This is why vector quantization is suitable for lossy data compression.

4.5 Time Normalization

Temperament, situation, time duration, location and environment always matter for a speaker. Thus, same word spoken may exhibit variations. One of the major variations is present in the duration of words. Problem starts for a pattern matcher who is required to compute the distance between an unknown word and a set of word templates. It is not possible to align both the beginning and the end of the word with the templates. One way to solve this problem is linear time warping algorithm that is to expand and compress the duration of each word so that it is equal to some fixed length such as the average length of the templates. It performs on small vocabulary over isolated word recognition system. This performance is not good for large vocabulary. The more general solution is nonlinear time normalization which is achieved using the dynamic programming and called as Dynamic Time Warping (Yadav, Munshi and Afshar Alam, 2010).

4.6 Dynamic Time Warping (Dtw) Algorithm

Time series data are available everywhere in the form of data occurrence in every scientific discipline. The common task of time series data is to compare data of one sequence to other (Tsinaslanidis, P. E., & Kugiumtzis, D., 2014). Euclidean distance method is useful when simple distance measures are required. However it is found that the two sequences are having approximately similar component but they cannot line up any axis. For finding the similarities between such types of sequences concept of warp time axis is useful to measure the similarity between patterns (T. Rakthanmanon, 2012). DTW algorithm helps to find the solution of such types of problems rendering Keogh, Eamonn J. and Michael J. Pazzani., (2001).

DTW algorithm is optimum dynamic programming approach (Richard Bellman, 1957) for spoken word recognition in speech (Sakoe, H. & Chiba, S., 1978). Furutuna Titus Felix (2008) proposed an alternative way to implement DTW algorithm for similarities measures between two unequal sequences in speech recognition. This method is more efficient than the existing algorithms. The existing DTW algorithms have computation time complexity $O(n^2v)$ for v number of words stored in the database (Wang, 2012; Yadav, Munshi and Afshar Alam, 2014).

Cuturi, Marco (2011) proposed a novel technique which cost the DTW distances and similarities as positive definite kernels in time sequences. A theoretical concept is developed in the family of global alignment kernels. The proposed alternative kernels are efficient in computing as well as are positive definite. The experimental results proved that these alternative techniques are more efficient as compared to others available kernels which are based on the DTW formalism.

Senin Pavel (2008) proposed the concept of analysis of software metric using DTW algorithm and introduced the naive DTW for software development in telemetry data. Using local and global parameters DTW speeds up through the scaling. DTW may also be used for similar query in the long streams for the software development in telemetry data. Another approach for double stage DTW algorithm using double stage in template matching for images (Somya Adwan and Hamzah Arof, 2010).

Lama, Palden and Mounika Namburu (2010) presented the technological advances in past few decades in the field of speech recognition system. Speaker independent algorithm is developed and implemented. In this algorithm extraction of salient features from the input speech signal is done and same is used for the isolated word as well as for connected word. The technique has been verified for voice to text convertor application of speech recognition in speaker dependent model.

Muda Lindasalwa et al., (2010) worked over viability of MFCC to extract the feature vectors from the speech and used DTW for pattern matching in speech recognition which produced better result. DTW based techniques are useful data mining applications for pattern recognition as well as in finance rendering Berndt and Clifford, (1994), Keogh Eamonn J. and Michael J. Pazzani, 2000 & 2001 and Tsinaslanidis et al., (2014).

Tsinaslanidis and Kugiumtzis (2014) suggested few key points to solve the problem of segment price series in business market as well as predication of the market. An algorithm is proposed for efficient market hypothesis (EMH) with two well-known data mining tools namely- perceptually important points (PIP) and DTW. Table 5 summarizes DTW algorithms.

Table 5: Summarizes DTW Algorithm

Authors	Highlights
Sakoe, H. & Chiba, S., -1978	Optimum dynamic programming approach for spoken word recognition in speech using the concept of dynamic programming
Keogh, Eamonn J., and Michael J. Pazzani, 2001	Two sequences having approximately similar component but they cannot line up any axis. For finding the similarity between such types of sequences the concept of warp the time axis used for getting better similarity measure.

Cuturi, Marco, 2011	A theoretical concept in the family of global alignment kernels (Cuturi et al. 2007). An alternative kernel was proposed which is efficient in computing as well as are positive definite.
FURUTUNA Titus Felix Titus., 2008	It is more efficient than the existing algorithms. The existing DTW algorithm having order of time complexity is O (n^2v) for v number of words which storage in the database.
Senin Pavel, 2008	The naive DTW uses local and global parameters for software development in telemetry data. In this technique DTW speeds up the system through scaling.
Somya Adwan and Hamzah Arof, 2010	A novel approach for double stage DTW algorithm using double stage in template matching for images.
Lama. Palden and Mounika Namburu, 2010	Technique is verified for voice to text convertor application of speech recognition in speaker dependent model.
Muda Lindasalwa et al., 2010	The viability of MFCC to extract the feature vectors from the speech and used DTW for pattern matching in speech recognition which produced better result.

4.7 Statistical Modelling

There are variations in the productions of words. A single production of a word generates a sequence of feature vectors such as spectra or frames of linear prediction coefficient. Other production of the same word generates similar but different sequences. The underlying model can be thought of as a sequence of states representing the feature vectors and the transmission between two states.

V. CONCLUSIONS

We presented in details about speech recognition, speech recognition in noisy environment, feature extraction techniques in speech recognition, pattern matching techniques and DTW algorithm in speech recognition.

VI. ACKNOWLEDGEMENT

We would like to express our sincere thanks to Dr. Abdul Mobin, Ex. Chief Scientist, National Physical Laboratory, Delhi, for his valuable guidance and sincere advise without which it could not be possible to complete the work presented in this paper.

VII. REFERENCES

[1]. Al-Haddad, Syed Abdul Rahman, Salina Abdul Samad, Aini Hussain, Khairul Anuar Ishak, and Hamid Mirvaziri. "Decision fusion for isolated Malay digit recognition using dynamic time warping (DTW) and Hidden Markov Model (HMM)" in Research and Development, SCOReD 2007. 5th Student Conference on, pp. 1-6. IEEE, 2007.

[2]. Ali, H., Ahmad, N., Zhou, X., Iqbal, K., & Ali, S. M., DWT features performance analysis for automatic speech recognition of Urdu. Springer Plus, 3(1), 204. 2014.

[3]. Azmi, Mohamed Mostafa, Hesham Tolba, Sherif Mahdy, and Mervat Fashal. "Syllable-based automatic Arabic speech recognition." In Proceedings of the 7th WSEAS International Conference on Signal Processing, Robotics and Automation, pp. 246-250. World Scientific and Engineering Academy and Society (WSEAS), 2008.

[4]. B. A. Dautrich, L. R. Rabiner, T.B. Martin, On the Use of Filter Bank Features for Isolated Word Recognition, ICASSP 83, BOSTON, IEEE 1983, pp. 1061-1064.

[5]. Bagul, S. G., and R. K. Shastri. "Text independent speaker recognition system using gmm." In Human Computer Interactions (ICHCI), 2013 International Conference on, pp. 1-5. IEEE, 2013.

[6]. Behroozmand, R., & Almasganj, F. Comparison of neural networks and support vector machines applied to optimized features extracted from patients' speech signal for classification of vocal fold inflammation. In Signal Processing and Information Technology, Proceedings of the Fifth IEEE International Symposium on (pp. 844-849). 2005.

[7]. Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time series. In KDD Workshop Vol. 10, No. 16, pp. 359-370.

[8]. Bhattacharjee, U., & Sarmah, K. (2013) Language identification system using MFCC and prosodic features. In Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on (pp. 194-197).

[9]. Bishop, Christopher M., Neural Networks for pattern recognition. Oxford University Press, 1995.

[10]. Burton, D., J. Shore, and J. Buck. "A generalization of isolated word recognition using vector quantization." In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83., vol. 8, pp. 1021-1024. IEEE, 1983.

[11]. Choi, E., & Lee, C. (2001). Optimizing feature extraction for multiclass problems. IEEE transactions on geoscience and remote sensing, 39(3), 521-528.

[12]. Choi, Euisun, and Chulhee Lee. "Feature extraction based on the Bhattacharyya distance." Pattern Recognition 36, no. 8 (2003): 1703-1709.

[13]. Choi, Euisun, and Chulhee Lee. "Feature extraction based on the Bhattacharyya distance." In Geoscience and Remote Sensing Symposium, 2000. Proceedings. IGARSS 2000. International, vol. 5, pp. 2146-2148. IEEE, 2000.

[14]. Cox, R. V., Kamm, C. A., Rabiner, L. R., Schroeter, J., & Wilpon, J. G. (2000). Speech and language processing for next-millennium communications services. Proceedings of the IEEE, 88(8), 1314-1337.

[15]. Cuturi, Marco. "Fast global alignment kernels." In Proceedings of the 28th international conference on machine learning (ICML-11), pp. 929-936. 2011.

[16]. Davis, K. H., R. Biddulph, and Stephen Balashek. "Automatic recognition of spoken digits." The Journal of the Acoustical Society of America 24, no. 6 (1952): 637-642.

[17]. De Poli, G., & Mion, L. (2006). From audio to content. Livro não publicado. Padova: Dipartimento di Ingegneria Dell'Informazione-Università degli Studi di Padova.

[18]. Drenthen, G. S. Speech Recognition using a Dynamic Time Wrapping approach. 2012. pp 1-6.

[19]. Duda, R. O., Hart, P. E., & Stork, D. G. Pattern classification. John Wiley & Sons. 2012.

[20]. Fleissner, Sebastian, Xiaoyue Liu, and Alex Fang. "Acoustic classification and speech recognition histories for adaptable spoken language dialogue systems." In Proceedings of the 17th International Congress of Phonetic Sciences, pp. 679-682. 2011.

[21]. Furtuna, Titus Felix. (2008), Dynamic programming algorithms in speech recognition. Revista Informatica Economică nr 2, no. 46, 94.

[22]. Gaafar, T. S., Bakr, H. M. A., & Abdalla, M. I. (2014, May). An improved method for speech/speaker recognition. In Informatics, Electronics & Vision (ICIEV), 2014 International Conference on (pp. 1-5). IEEE.

[23]. Gupta, Kshitij, and John D. Owens. "Compute & memory optimizations for high-quality speech recognition on low-end GPU processors." In High Performance Computing (HiPC), 2011 18th International Conference on, pp. 1-10. IEEE, 2011.

[24]. Hansen, Per K. "Voice control system." U.S. Patent 4,776,016, issued October 4, 1988.

[25]. Hayasaka, Noboru, Kham Khankhavivone, Yoshikazu Miyanaga, and Kraisin Songwatana. "New robust speech recognition by using nonlinear running spectrurn filter." In Communications and Information Technologies, 2006. ISCIT'06. International Symposium on, pp. 133-136. IEEE, 2006.

[26]. Hongbin, Gao, Pang Weiyi, Huang Chunru, and Zhang Yongqiang. "A speech endpoint detection based on dynamically updated threshold of box-counting dimension." In Information Technology and Applications, 2009. IFITA'09. International Forum on, vol. 2, pp. 397-401. IEEE, 2009.

[27]. Jiang, Hai, Meng Joo Er, and Yang Gao. "Feature extraction using wavelet packets strategy." In Decision and Control, 2003. Proceedings. 42nd IEEE Conference on, vol. 5, pp. 4517-4520. IEEE, 2003.

[28]. Kalamani, M., S. Valarmathy, C. Poonkuzhali, and J. N. Catherine. "Feature selection algorithms for automatic speech recognition." In Computer Communication and Informatics (ICCCI), 2014 International Conference on, pp. 1-7. IEEE, 2014.

[29]. Kang, Guangyu, and Shize Guo. "Variable sliding window DTW speech identification algorithm." In Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on, vol. 1, pp. 304-307. IEEE, 2009.

[30]. Keogh, Eamonn J., and Michael J. Pazzani. "Derivative dynamic time warping." In Proceedings of the 2001 SIAM International Conference on Data Mining, pp. 1-11. Society for Industrial and Applied Mathematics, 2001.

[31]. Keogh, Eamonn J., and Michael J. Pazzani. "Scaling up dynamic time warping for data mining applications." In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 285-289. ACM, 2000.

[32]. Lahouti, Farshad, Ahmad R. Fazel, Amir H. Safavi-Naeini, and Amir K. Khandani. "Single and double frame coding of speech LPC parameters using a lattice-based quantization scheme." IEEE Transactions on Audio, Speech, and Language Processing 14, no. 5 (2006): 1624-1632.

[33]. Lama, Palden, and Mounika Namburu. Speech recognition with dynamic time warping using MATLAB. Project Report, CS 525, Springer 2010.

[34]. Lee, Chulhee, and Euisun Choi "Optimizing feature extraction for multiclass problems." In Pattern Recognition, 2000. Proceedings. 15th International Conference on, vol. 2, pp. 402-405. IEEE, 2000.

[35]. Lee, Chulhee, Euisun Choi, and Jaehong Kim. "Optimal feature extraction for normally distributed data." In Aerospace/Defense Sensing and Controls, pp. 223-232. International Society for Optics and Photonics, 1998.

[36]. Midorikawa, Yoichi, Yuta Muraoka, and Masanori Akita. "Noisy speech recognition using wavelet transform and weighting coefficients for a specific level." In Proceedings of 20th International Congress on Acoustics, Sydney, Australia, pp. 1-7. 2010.

[37]. Miyanaga, Y., Takahashi, W., & Yoshizawa, S. (2013). A Robust Speech Communication into Smart Info-Media System. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 96(11), 2074-2080.

[38]. Mobin, A., Agrawal, S. S., Ganesan, M., & Pavate, K. D. (1989, November). A novel technique of data compression for spoken word recognition systems. In TENCON'89. Fourth IEEE Region 10 International Conference (pp. 756-759). IEEE.

[39]. Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv: 1003.4083 (2010).

[40]. Nehe, N. S., and R. S. Holambe. "New feature extraction methods using DWT and LPC for isolated word recognition." In TENCON 2008-2008 IEEE Region 10 Conference, pp. 1-6. IEEE, 2008.

[41]. Nica, Alina, Alexandru Caruntu, Gavril Toderean, and Ovidiu Buza. "Analysis and synthesis of vowels using Matlab." In Automation, Quality and Testing, Robotics, 2006 IEEE International Conference on, vol. 2, pp. 371-374. IEEE, 2006.

[42]. Niemann, H., Lang, M., & Sagerer, G. (Eds.). (2012). Recent advances in speech understanding and dialog systems (Vol. 46). Springer Science & Business Media.

[43]. Pai, H. F., & Wang, H. C. (1993). A two-dimensional cepstrum approach for the recognition of Mandarin syllable initials. Pattern recognition, 26(4), 569-577.

[44]. R. E. Bellman, Dynamic Programming, Princeton University Press, Princeton, New Jersey, USA, 1957.

[45]. Rabiner, L. R., & Juang, B.-H. (2006). Speech recognition: Statistical methods. In K. Brown (Ed.), Encyclopedia of Language & Linguistics (pp. 1-18). Amsterdam: Elsevier. doi:10.1016/B0-08-044854-2/00907-X

[46]. Rahman, M. M., Khan, M. F., & Moni, M. A. (2010). Speech recognition front-end for segmenting and clustering continuous Bangla speech. Daffodil International University Journal of Science and Technology, 5(1), 67-72.

[47]. Rice, S. O. Mathematical Analysis of Random Noise, Bell System Technical Journal, July, 1944, Vol.23, pp-282-332, January 1945, Vol. 24, pp 46-156.

[48]. Sakoe, H., & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing, 26(1), pp-43-49, 1978.

[49]. Schroeder, Manfred R. "Linear prediction, extermal entropy and prior information in speech signal analysis and synthesis." Speech Communication 1, No. 1: pp 9-20, 1982.

[50]. Schwartz, Mischa. "Improving the noise performance of communication systems: 1930s to early 1940s." In History of Telecommunications Conference, 2008. HISTELCON 2008. IEEE, pp. 72-78. IEEE, 2008.
[51]. Senin, Pavel. "Dynamic time warping algorithm review." Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA 855 (2008): pp. 1-23.
[52]. Shannon, C. E., A Mathematical Theory of Communication, Bell System Technical Journal, July, 1948, Vol. 27, pp 379-423.
[53]. Shannon, Claude E. "Communication theory of secrecy systems." Bell Labs Technical Journal 28, no. 4 (1949): 656-715.
[54]. Somya Adwan and Hamzah Arof, A Novel Double Stage Dynamic Time Warping Algorithm for Image Template Matching, 6th IMT-GT Conference on Mathematics. Statistics and its Applications (ICMSA 2010), University Tunku Abdul Rahman, Kuala Lumper, Malaysa. pp 667-667-676.
[55]. Srisai, D., & Ratanamahatana, C. A. (2009, November). Efficient time series classification under template matching using time warping alignment. In Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on (pp. 685-690). IEEE.
[56]. T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12, 2012, pp. 262.
[57]. Tsinaslanidis, Prodromos, Antonis Alexandridis, Achilleas Zapranis, and Efstratios Livanis. "Dynamic time warping as a similarity measure: applications in finance." no. Journal Article (2014).
[58]. W. A. Ainswirth, Speech Recognition by Machine, Vol. 12 of IEE Computing Series, Peter Peregrinus Ltd., London 1988.
[59]. W. A. Ainsworth, Some approaches to automatic speech recognition. In W. J. Hardcastle & J. Laver (Eds.), the Handbook of Phonetic Sciences. Oxford, Blackwell. 1997. pp. 721-743.
[60]. Wang, Gang-Jin, Chi Xie, Feng Han, and Bo Sun. "Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence from minimal spanning tree." Physica A: Statistical Mechanics and its Applications 391, no. 16 (2012): 4136-4146.
[61]. Xie, Lei, and Zhi-Qiang Liu. "A comparative study of audio features for audio-to-visual conversion in mpeg-4 compliant facial animation." In Machine Learning and Cybernetics, International Conference on, pp. 4359-4364. IEEE, 2006.
[62]. Xu, Gang, Bo Tong, and XiaoWei He. "Robust endpoint detection in Mandarin based on MFCC and short-time correlation coefficient." In Intelligent Computation Technology and Automation, 2009. ICICTA'09. Second International Conference on, vol. 2, pp. 336-339. IEEE, 2009.
[63]. Yadav, Munshi and Afshar Aalam. "Five Stage Dynamic Time Warping Algorithm for Speaker Dependent Isolated Word Recognition in Speech." International Journal of Computer and Software Engineering (IJCSE), (2016). Volume-4, Issue-10, pp. 112-115.
[64]. Yadav, Munshi and Afshar Alam. "Reduction of Computation Time in Pattern Matching for Speech Recognition." International Journal of Computer Applications 90, No. 18 (2014). pp. 35-37.
[65]. Yadav, Munshi and Afshar Alam. "A Novel Method for High performance Computing in Speech Recognition", International Conference on Next Generation Communication and Computing Systems (ICNGC2S-10), December 25-26, 2010, Chandigarh, India. pp. 345 - 347.
[66]. Zhu, Q., N. Ohtsuki, Y. Miyanaga, and N. Yoshida. "Robust speech analysis in noisy environment using running spectrum filtering." In Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on, vol. 2, pp. 995-1000. IEEE, 2004.
[67]. Zue, Victor, James Glass, Michael Phillips, and Stephanie Seneff. "The MIT SUMMIT speech recognition system: A progress report." In Proceedings of the workshop on Speech and Natural Language, pp. 179-189.Association for Computational Linguistics, 1989.
[68]. Zunkler, Klaus, An ISDN speech server based on speaker independent continuous Hidden Markov Models, NATO ASI Series, Vol. F 75, 1991.

Author is Graduated in Electronics and Communication Engineering from Govind Ballbh Pant Engineering College, Pauri Garhwal, Uttrakhand, University of H. N. Bahuguna Garhal Central University, Srinagar Garhwal, Uttrakhand. Post Grudated in Information Technology from University School of Information Technology, Guru Govind Singh Indraprastha Univesrsity, Delhi. Purusing PhD in Computer Science & Engineering from School of Engineering Sciences and Technology, Jamia Hamdard, Delhi