1. INTRODUCTION

Pattern recognition is the process of classification of data using the knowledge already gained or with the help of statistical information derived from patterns or the representation of data. Speech Recognition is one of the popular pattern recognition applications that have become more significant as its scope has extended to most of the important arenas involved to lead day-to-day life. Most prominent domains that involves speech recognition are health care, speech assisted gadgets for disables, in-car systems, military, telephony, court reporting, multimodal interaction, hands-free computing, telematics, video games, virtual assistant and more. The growth and availability of advancements in digital hardware at affordable cost has tremendously supported the growth of computational engineering, artificial intelligence, pattern recognition and machine learning leading to solve more and more complex problems. This thesis titled "Graph based Segmentation and Deep Learning for Phoneme Pattern Classification in Tamil Continuous Speech" is focused on the design and implementation of computational framework and deep learning models to develop phoneme pattern classification systems for phonemes in Tamil continuous speech. This chapter gives an introduction to pattern recognition, speech recognition, pattern recognition approaches to speech recognition and types of speech recognition systems. It then presents the review of literature and the objectives of the research.

1.1 PATTERN RECOGNITION

Pattern recognition is nothing but automated identification of patterns and regularities in data with the help of artificial intelligence or machine learning algorithms. The process of discovering patterns from the instances based on statistical information is called as pattern recognition. One can build highly potential applications to handle complex problems using pattern recognition.

Pattern recognition applications generally take the raw data, process and convert them into machine amenable form. Pattern recognition can be applied to solve both classification and clustering problems. Classification uses a supervised learning approach which classifies the incoming data to a class label where as clustering uses unsupervised approach where the training data is clustered based on their patterns and forms the decision rules that enables to identify a new incoming data. Before performing any pattern recognition tasks, the features are extracted from raw data through one of the feature extraction technique and are represented as feature vectors for further learning. Some of the successful applications that uncovered the potential benefits of pattern recognition are speech recognition, automatic medical diagnosis, speaker

identification, Multimedia Document Recognition (MDR), image processing, computer vision, seismic analysis, radar signal classification or analysis, finger print identification and are elucidated below except speech recognition which is detailed later in this chapter.

There are several problems in medical diagnosis that uses pattern recognition to automate or assist the diagnosis process [1]. Multivariate time series data from clinical instruments, images captured by the clinical instruments like MRI, CT scan, X-ray, etc. are used in the pattern recognition approaches to build decision support systems for supporting the diagnostic process. For example, pattern recognition approaches are applied in sequence learning problems, processing and analyzing output signals from electrocardiograms, glucose meter and in genomics. Further, several deep learning techniques are applied to analyze the psychological conditions, head injuries, Parkinson disease, cancer detection, lung infections, etc. Machine learning approach like neural networks are used for artifact removal in clinical data, for early detection and prediction of targeted abnormalities like cancer affected cells, for clustering and subtyping infected regions, for modelling bradycardias like physiological events and more.

Computer vision is an application area which is filled up with huge set of problems including gesture recognition, object recognition, scene recognition, action recognition, computer vision for autonomous vehicle, iris recognition, etc. Gesture recognition uses pattern recognition to track the gesture of the user in front of the system, mapping it to the corresponding representation and transforming to a command. The objects in images or videos are identified and labelled in object recognition using pattern classification approaches like template matching after the features of the intended objects are extracted. Computer vision plays a major role in processing the input from the computerised vision of an autonomous vehicle. Here, the real-time video captured from the environment is then processed to detect obstacles, assist in navigation and to act environment responsive using powerful machine learning approaches to pattern recognition like deep learning.

Speaker identification is another application domain of pattern recognition which is to identify the speaker belonging to the speech uttered, from the registered set of people. Speaker recognition systems can be built on either an open set or closed set of speakers, can be text dependent or independent or can be applied for speaker identification or verification process. In these systems, once the database with required samples is available, the features are extracted from samples of various speakers and passed to a train the pattern recognition model like neural networks, support vector machines, hidden markov model, etc. Various types of features are available that supports in speaker identification which includes spectral features to extract the

physical characteristics of vocal tract, dynamic features to extract the time evolution in the spectra, source features to extract the glottal characteristics of voice, suprasegmental features to extract characteristics across different segments of the speech and high level features to define the symbolic characteristics of a speaker.

Seismic analysis is another area where pattern recognition is applied and helps to save lives by enabling seismologist in estimating the shaking hazards and motion parameters at a site under consideration. It helps to propose designs resistant to earthquake and take precautions with seismic safety assessment. The seismic analysis is even extended to the field of oil exploration and also helps in risk analysis and hazard analysis. The pattern recognition algorithms are used to help identify the earthquake prone areas that react greater than a specified threshold which is defined safe. The pattern recognition approaches assist in analyzing and identifying the parameters that contribute enough in earthquake. Few of such identified parameters are intersection on lineaments, intersection faults, etc. helps in evaluating the stress accumulation at a location. CORA-3 is a popular pattern recognition algorithm that is used in various studies to identify the earthquake prone areas at various regions.

Radar signal classification and analysis is another domain where the foot prints of pattern recognition play a major role [2]. Earlier, they were used only in military for surveillance, weapon guidance and navigation but now-a-days they are used even for civilian tasks like traffic control, navigation, pollution control, space observation, weather forecast, etc. These tasks are implemented by performing various subtasks like detecting, identifying, locating and tracing objects around. These systems receives the signals from the radar, sends to the signal processors for extracting the signal features respective to the problem, which are then processed and submitted to the pattern recognition models to perform the recognition or detection tasks. Electronic support measures provide reliable information about the radiated electromagnetic energy that is used in threat detection, avoidance and real-time counter measures. Most systems related to radar emitter recognition and identification use neural networks due to its powerful parallel architecture that is capable of handling even incomplete data.

Pattern recognition follows a sequence of subtask to solve any pattern recognition problem which is discussed below.

Pattern Recognition Process

The investigation of a problem in the pattern recognition process involves several steps which are listed out below [3]:

1. *Formulation of the problem* - Understanding clearly about the objective of the problem under investigation and planning the other stages. This involves an explorative search to clarify the problem considered. This is an iterative process which involves exploration and problem refining that ends up with planning.
2. *Data collection* - Collecting the required data based on the ground truth of the problem and recording the procedure of data collection.
3. *Initial examination of the data* - Verifying the data, evaluating the summary statistics and creating plots in order to reveal its structure.
4. *Feature selection or feature extraction* - Selecting the features of the measured data available in the data set that are appropriate solving the task under consideration. These new features can be obtained either by a linear or nonlinear transformation of the original data set, which is termed as feature extraction. The dissection of feature extraction and classification is artificial to some extent.
5. *Unsupervised pattern classification or clustering* - This phase performs an exploratory data analysis and directs the investigation to provide a successful conclusion. Alternatively, it may act as a means of pre-processing the data for a supervised classification process.
6. *Apply appropriate discrimination or regression procedures* - The classifier is developed with training set of exemplar patterns.
7. *Assessment of results* - This involves the application of the trained classifier to a test set of labelled patterns that helps in evaluation of the classifier constructed.
8. *Interpretation* - Transforming the results to information as required for presenting a solution to the problem.

Necessarily, the above discussed steps are requirded to be implemented as iterative processes that may expand the hypothesis further, thus requiring huge collection of data to train the models.

Paattern Recognition Models

Several approaches have been developed for solving pattern recognition problems. Based on the method of data mining and classification applied, the pattern recognition models can be classified as statistical model, structural model, template matching, fuzzy-based, neural-network based, machine learning models, etc.

- ***Statistical Model*** **-** This method mostly defines the probability distributions describing the discriminating features of various classes [4]. Thus, it can be defined that these models use statistical basis for data classification. Bayesian classification, Correlation and statistical hypothesis testing are examples of statistical models. The ability to represent patterns defining the respective classes to separate itself from other classes drives the effectiveness of the model. Generally, the error probability is used to evaluate the nearness of a new input sample to a class. Sometimes, discriminate analysis based approach is used by defining functions like linear or quadratic to represent the decision boundary in a parametric form to help in the process of classification [5] when the problem complexity is high or the dimensionality of the system is high.
- ***Structural Model*** **-** Unlike statistical approach, this can be used for complex problems, where the patterns directly could not be discriminated. They use simpler sub-patterns called primitives and their interrelationships to identify the patterns embedded on the data points to discriminate its class. This approach helps in yielding a combinatorial explosion of maximum possibility to be investigated. This requires large training samples and in turn more computing resources [6]. In this approach, the patterns are represented as a set of structures defined in the form of a tree, graph, formal language or string. Then the unknown pattern is represented as a structure which is generated automatically and recognized using some parsers.
- ***Template Matching Model*** **–** The matching process is done by finding the similarity between two entities of same type, for example points, curves or shapes [7]. The pattern of each class is defined as a prototype. The pattern to be classified is then compared against prototypes of various classes using a similarity measure like correlation. These models generally are resistant to translation, rotation or scaling operations. Eventhough the models works good for the classification problems in most of the domains, it has the disadvantage to handle distorted data if its deformation process is indirect. Deformation is required when the patterns could not be directly designed, in that situation the deformable template matching could be implemented.

- ***Neural Network Based Model*** **-** A self-adaptive network that comprises of massive simple processing units termed as neurons [8]. The neurons are interconnected with a high degree where they work cooperatively with each other, thus achieving massive parallel distributed processing. The design of the neural network imitates the biological brains where as its functionality resembles the neural systems. The application of neural network has shown its success for a huge set of problems in machine learning and pattern recognition. The neural network has evolved with a collection of algorithms that supports in the learning process of the network enabling the synaptic connections between the neuron to evolve pertaining to a specific task. They have the ability to learn the interrelationship between complex input output data through sequential procedures in a self adaptive manner. Neural network are popularly applied even in feature extraction and classification problems. It is also possible to map other feature extraction and classification methods to neural network to improve the method efficiently specially in complex non-linear problems. Eventhough the neural networks have their own underlying framework and principles most of the popular neural network models, they are by implicit equivalent to a typical statistical pattern recognition model.

- ***Fuzzy Based Model*** **-** These models use the concept of fuzzy sets/fuzzy logic/soft labelling to define the classes building the model [9]. This type of pattern recognition model is based on the degree of membership otherwise called as soft labels of the data to a particular class rather than enforcing mutual exclusion constraint between the competing classes as seen in other types. It uses the membership values of the unseen data to the classes in deciding the output. Unlike classical pattern recognition, these models are transparently designed making the logic and steps undergone in the decision making process traceable. These models can be built either with the help of expert's knowledge or data or both. Fuzzy based models can be subtyped as either rule-based or prototype based. The Fuzzy rule based model build a set of rules to define the system to predict the class of unknown input data, where as fuzzy prototype based model applies fuzzifies the definitions of classes in classical pattern recognition models like k- nearest neighbour (K-nn). For example the fuzzy K-nn uses distance measure along with the soft labels to predict the class of the unknown data.

- ***Hybrid Model*** **-** These models integrate the functionality of two or more models discussed above. The limitations observed in the models discussed above are overcome by inheriting the pros of two or more models to form a hybrid model. Some popular hybrid models are Neuro-fuzzy models, Support Vector Machine – Hidden Markov Model (SVM-HMM), Neural Network-Hidden Markov Model (NN-HMM), Genetic algorithms and decision trees, etc. [10

– 13]. For example, a Neuro-Fuzzy model takes advantage of (i) the reasoning capability of fuzzy systems that uses fuzzy sets and linguistic models, (ii) the universal approximation capability of neural networks making these systems with higher interpretability and accuracy. At the same time, in a SVM-HMM model, the discriminating capabilities of SVMs are incorporated to perform the classification while the temporal details on the data are structured and controlled by HMM.

Supervised Vs Unsupervised Models - Supervised learning is an approach to model a system where the input variables, x and an output variable, y are given and an algorithm is used to map the input data to the output. It can be defined as, $y = f(x)$. The basic objective here is to develop an optimized mapping function, so that it predicts the output with maximum accuracy for a given new input. It is called supervised learning, as the learning process is entirely monitored and controlled by the training dataset that acts as a teacher guiding their student. The learning process is iterative, that allows it to improve its performance by learning more and more in each iteration, and continues so till it reaches a desired level of accuracy or performance.

In contrast to supervised learning, unsupervised learning is an approach that learns from the input, x, alone. The objective of unsupervised learning is to learn more on the data by modelling the distribution in the data. It is called unsupervised, as the model building process is not guided by any output variable, i.e. learns without a teacher. The algorithms take the responsibility of discovering the hidden structures available in the data to model the system.

1.2 SPEECH RECOGNITION SYSTEMS

Speech is the most powerful and widely used communication mode in the society. Speech recognition is nothing but making machines understand and react to the human speech. Speech recognition aids as an effective alternate interface to interact with the machine even for illiterate and disabled community. The world today is filled with a variety of gadgets and machines around, each servicing the community in some way or the other. Most of those devices available now-a-days are seen with the voice or speech enabled capability. Speech technology helps a lot to transform the environment into a smart environment by making the machines and appliances around us to respond our commands. Before interpreting any speech, the microphone in the machines needs to translate the voice vibrations of a person's speech into electrical signals, where some hardware like sound card converts it into digital signals. These signals are then analyzed by the speech recognition systems to recognize the sentences, words, syllables or phonemes spoken. Phonemes are the basic building blocks of speech. The phonemes can then be

recombined to form words which in turn are combined to form sentences. Sometimes, two or more words sound alike which are further recognized based on their context under consideration. Based on the complexity level of speech uttered, speech recognition can be classified into isolated word recognition, connected words recognition, continuous speech recognition and spontaneous speech recognition. Blockbusters of information technology like Amazon, Google, Apple, Microsoft, etc. have come out with low cost commercial applications for interactive speech recognition [14].

1.2.1 Pattern Recognition Approaches for Speech Recognition

History of speech recognition reveal that pattern recognition is a promising methodology in building good and competitive speech recognition models. With the knowledge of various pattern recognition models available as discussed in the previous section, here various approaches in the area of speech recognition are discussed.

Template Based Speech Recognition

This approach builds a collection of prototypes, each representing a word in the dictionary. Once the system is built, a new utterance given to the system is matched with all the prototypes in the collection, to choose the best matching pattern. Generally, templates are constructed for whole words, thus much suitable for small to medium vocabulary word recognition systems and connectionist systems, eliminating the errors caused due to acoustically variable units like phonemes [15, 16]. This approach has the advantage of building perfect word models with the drawback of holding fixed pre-recorded templates. The variability in the properties of speech by different speaker adds complexity to the system, which can be solved by adding multiple templates per word. But this works fine for dictionaries less than 100 words, otherwise making the system inefficient in terms of memory and processing power. Also the approach is not suitable for continuous speech recognition where the speech variability is high.

Knowledge Based Speech Recognition

This approach to speech recognition uses the expert knowledge on linguistic, phonetic and spectrogram of speech to build the production rules of the system. But building such production rules is a tedious task and impractical for speech due to its high variability. So, techniques like decision trees are used to select appropriate features from the feature set that support in the process of classification and generate the production rule set to build the knowledge base. The inference engine of the system is responsible for identifying the appropriate rule with the high

firing strength to classify the given input speech. These approaches can best be used for incorporating the expert's knowledge to the speech recognition system [17-19].

Neural Network Based Speech Recognition

Basically, neural networks are network of neuron mimicking the biological neurons in brain and their functioning. These neurons in Artificial Neural Networks (ANNs) are trained in parallel. The energy in various frequencies of the speech signal do occur in parallel, where the subunits like phonemes or syllables occur in serial. The frequencies at various time units of speech are given as input to the neural network to train the neurons in the network to model the subunits of the speech. The neurons are connected with weighted links in the network controlling the weightage of a particular input to a neuron, where the neurons hold an activation function that produces an output for the given set of inputs. The main challenge here lies in identifying the appropriate connection weights and parameters defining the activation functions. Popular techniques like back propagation, gradient-decent help in training the ANNs. Once the model is trained with the training subword units, it becomes ready to classify the unseen subword speech units.

Dynamic Time Warping (DTW) Based Speech Recognition

This pattern matching algorithm compares the input words with the reference words. Every reference word is denoted using a spectrum and does not show any distinction in between the subword units of the word. Different speakers pronounce the same word with different speed which can be matched with an algorithm like DTW that is capable of handling non-linear fluctuation occurring in the speech to time of various instances. The algorithm stretches or shrinks by warping the time axis of unknown word until a maximum coincidence with the word under reference is achieved [20].

Statistical Based Speech Recognition

The systems make use of mathematical functions and probability to discover the most likely output of a classifier. Hidden Markov Models (HMMs) are most powerful, successful and popularly used classifiers for speech recognition. The other is neural network which is currently gaining more attention in most hot areas with deep architectures. The HMMs models the spoken word as a link of phonemes, where the phonemes are represented as states of a HMM. During the searching process the chain tries to branch in different directions attempting to identify the most likely phoneme that comes next. The branching decisions here are supported by the probabilities of phonemes that are evaluated with the help of associated dictionary and training. In case of

sentences and phrases, process becomes more complicated as the system has to even identify the starts and stops of each word [21-23].

1.2.2 Classification of Speech Recognition System

Speech recognition systems are categorized based on various parameters like type of speech utterance given as input to the system, based on the vocabulary size used to train the system and based on the speaker mode. Various categorizations of speech recognition systems are mentioned below.

Classification Based on type of Speech Utterances

The speech utterances vary from person to person. Even the same person utters the speech differently at different situations. When concerned with the automatic speech recognition, the speech uttered by a person is conditioned on the application or the user interface module used for the speech recognition. Some system requires the user to recite one word at a time, some may allow few words at a stretch and some allow natural speech. With respect to the problem under consideration, the speech recognition systems can be classified as follows based on the speech utterances allowed by the system.

Isolated Word Recognition: These systems are built to recognize a word at a time. The input to this system is the speech signal of single word at a time. So, the system will be in either listen or non-listen states. The words under consideration need to be trained. These systems mostly rely on designing good acoustic models and least depends on language models. Consider that each word is represented as a sequence of observations $O = o_1, o_2, o_3, \ldots . o_T$ otherwise called as speech feature vectors observed at various time units, $t = 1\ to\ T$. Then the isolated words can be recognized by evaluating the $\arg max(P(w_i|O))$, with w_i denoting the i^{th} word in the vocabulary. The probability of the input word to be $w_i$cannot be defined directly but can be defined using the Baye's rule in terms of the prior probabilities, $P(w_i)$ and likelihood $P(O|w_i)$ and is given by,

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)} \tag{1.1}$$

Connected Words Recognition: These are like isolated word recognition, in addition, allowing speech utterances of sequence of words with enough pause between each word. Connected words recognition systems are generally built by extending the isolated word recognition system by including a boundary detection module to detect the boundaries of the words in connected speech signal and a language model that helps to model the correspondence lying between the

acoustic signals and words. The language models are generally built by incorporating the likelihood of neighbouring words. The recognition module in this case does not have the knowledge, if the word it receives is from a connected speech or itself is isolated.

Continuous Speech Recognition: These systems receive continuous speech as input where it does not have any knowledge about number of spoken words, phonetic units or segmentation details. The system has to identify the basic units of the speech, by detecting its boundaries and recognize the speech units. The objective of these systems is to come out with a sequence of speech units like phonetic segments, words or sentences for the given input speech signal. These systems face challenges like variability in linguistics, speaker and channel. This model is built with the underlying events available as a reference to build the structural and statistical model which is then used to recognize the events of the incoming input speech.

Spontaneous Speech Recognition: These systems have the capability to handle natural sounding speech that is not rehearsed. These systems have the ability to lever the challenges like additional sounds running along with the actual speech including "ums", "ahs" and even shutters which can be noticed naturally in common speech, duplicate utterance of words, partial words, filled pause and more. The system should extract the speech signals corresponding to the actual speech by removing the irrelevant parts in the speech signal before performing the recognition process.

Classification based on the Vocabulary Size

The size of the vocabulary considered while designing a speech recognition system plays a major role in defining the complexity of the model. The complexity of the system increases with the increase in vocabulary size, which in turn also increase the computing resources and adds more challenge in building an accurate system. So, the automatic speech recognition system based on the vocabulary size is categorized as small vocabulary system (10 words), medium vocabulary system (100 words), large vocabulary system (1000 words), very large vocabulary system (more than 10000 words) and out-of vocabulary system, that is able to map a word form vocabulary to unseen word.

Classification based on Mode of Speaker

The speech recognition module built may be either speaker dependent or independent based on the application of module. The classification of speech recognition systems based on speaker mode is done as speaker independent and speaker dependent systems.

Speaker Independent: The system is trained with the speech samples collected from a variety of speakers, enabling to build a system robust enough to handle speech of various speakers. The samples in the training set include speech from a collection of speakers covering all words in the vocabulary. The templates of words are also derived by cross section of speakers of various categories of age group, dialect, ascent and gender. The models are built by defining representative patterns for the words in the vocabulary.

Speaker Dependent: These systems are trained with the samples collected from predefined speaker. These systems improve the accuracy by adapting itself to the speaker. The complexity of these system are less when compare the speaker independent systems.

1.2.3 Continuous Speech Recognition Systems

A continuous speech recognition problem requires a model to be developed that accepts continuous speech from the speaker and is capable of recognizing a large collection of words. CSR extended to Large Vocabulary CSR (LVCSR) if the size of the vocabulary is more than 10,000. Sometimes, these systems should be capable of even identifying unseen words. Generally, the probabilistic modeling approach is used for large vocabulary speech recognition where specified word sequence produces a sequence of acoustic observations. To start with the continuous speech recognition, the input continuous speech need to be segmented into subword speech units. The subword speech units, then enters a recognition process, that is followed by a language modeling unit for completing the speech recognition. Building word models will yield better results, and will be suitable for isolated word recognition and connected word recognition systems, as in these cases, the properties of the words are well defined. But, in case of CSR there are two limitations, the training set need to be with enough instances of all words in the vocabulary to build a reliable model. For larger vocabulary and continuous speech the number of words is larger, the number of context of those words is much larger, thus increasing the complexity, making the process impractical. Secondly, the phonetic content in the individual words highly overlap in the context of large vocabulary as the constituent sounds are treated independently. This forces to develop a more efficient model of speech representation for continuous speech recognition.

At present, continuous speech recognition has reached its space in several application areas including health care and applications to assist the disabled people. The popular use of speech recognition in health care section is medical documentation. Speech recognition can be used in front-end or back-end of medical documentation process. In front-end speech recognition, the

dictated speech enter the speech recognition system where the recognized words are displayed on the screen and then the speaker is responsible for further editing and approval of the document. On the other hand, in the back-end speech recognition the speaker dictates to a digital dictation system, where the voice is directed to a speech recognition machine. The machine delivers the draft document along with the input voice file to the editor for further editing and approval. Currently back-end speech recognition, otherwise called as deferred speech is used widely in the industry.

As a part of radiology or pathology interpretation, discharge summary or progress note, speech recognition is most suited for the generation of narrative text. The ergonomic benefits in using speech recognition is comparatively minimal for people with sight and capable of operating the system through keyboard and mouse while concerned with entering structured discrete data like numbers, codes and controlled vocabulary. Apart from using speech recognition for document writing, it can be used as a therapy for brain arteriovenous malformation patients, where the prolonged use with word processors provides the benefit of re-strengthening short-term memory those have been treated with resection.

Lots of benefits can be realized through products with speech recognition as an interface by the people with disabilities. People facing challenges due to the hard of hearing problem or deaf can use these speech recognition based interfaces that help to automatically generate closed-captions for the speech conversation or discussion held in classrooms lectures, seminars, conferences, etc. These interfaces also help the people facing problems in using their hands temporarily or permanently for communicating with the gadgets or computers through keyboard or mouse. Speech recognition is used in products making use of captioned telephony, relay services, deaf telephony, etc. People who are not familiar or comfortable with text communication can get benefited out of this. This technology even support for the people with dyslexia but the effectiveness of the program lies in how effectively it is able to interpret the word spoken by the affected person.

1.2.4 General Framework for Phoneme Recognition in Continuous Speech

Automatic phoneme recognition is the process of mapping speech from a continuous-time signal to phonemes or speech sounds. The major challenge to achieve high-accuracy in recognition is the larger variability of speech signal characteristics. Different types of variability like linguistic variability, speaker variability and channel variability contribute to the variability of speech signal. Linguistic variability is caused by the effects of phonology, phonetics, syntax,

semantics and discourse on the speech signal. Speaker variability is caused due to intra/inter-speaker variability, which is as a result of co-articulation effect and due continuity and motion constraints on the human articulatory apparatus. Channel variability occurs due to the background noise and the hardware used in the transmission channel. The recognition process should be powerful enough to tackle all the aforesaid uncertainties to achieve good recognition accuracy. The general system for phoneme recognition in continuous speech is given in Fig. 1.1. The major steps involved in phoneme recognition are revealed in the forthcoming paragraphs.

Pre-processing Speech

In any real-time applications of speech, noise is an inevitable component of the speech signal. Microphone sensor in the communication channel captures the speech signal in addition to the environmental noise which leads to a distorted signal. The basic cleaning of the signal to improve its quality is generally done by filtering the speech as an initial process to improve the results of any speech processing techniques. There are several algorithms for performing noise reduction in speech. Adaptive noise cancellation and adaptive spectral subtraction are two popular methods used for noise reduction. Speech preprocessing can be done for various other purposes in addition to noise removal, which includes voice activity detection, pre-emphasis, framing and windowing. There are linear filter methods that perform time domain based filtering. Finitive Impulse Response (FIR) filters and Infinite Impulse Response (IIR) filters come under the category of linear filters. Apart from this, there are frequency domain based filters like low-pass, high-pass, band-pass and band-stop filters. Voice activity detection (VAD) is used in the preprocessing phase to identify the endpoints of the speech utterance which helps to improve the performance of the speech recognition systems. Various algorithms like zero crossing rate, energy, autocorrelation function based algorithms for VAD. Further, framing and windowing are the techniques used to split the speech signal into some fixed sized signals to enable further processing.

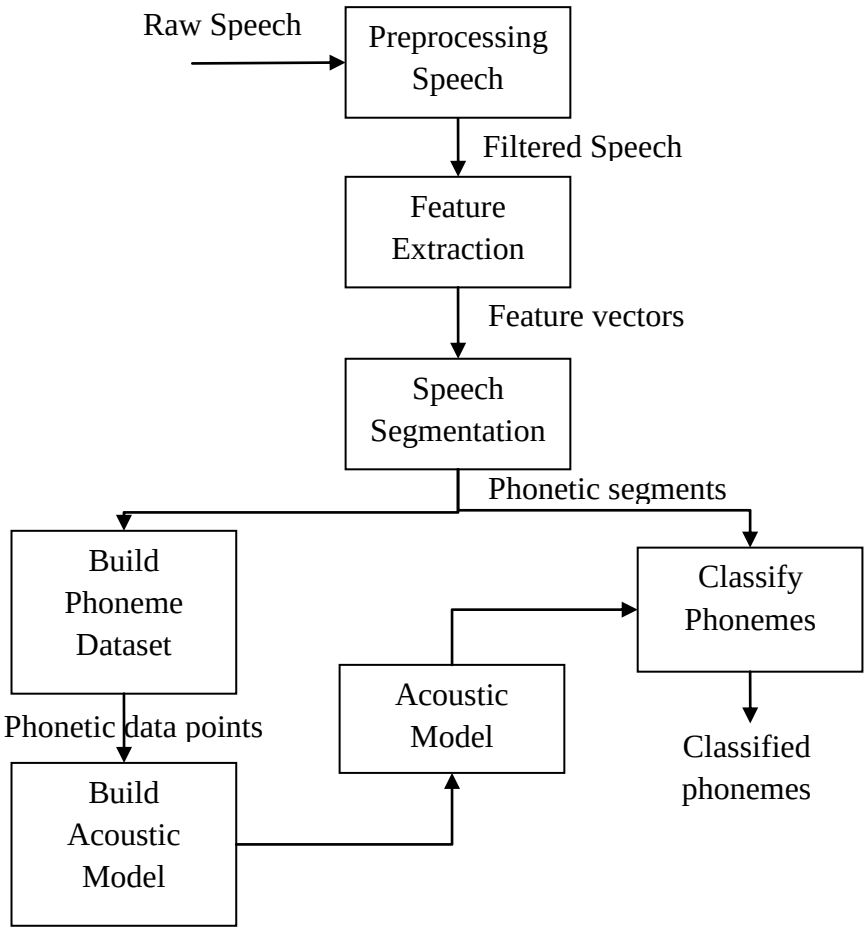

Fig. 1.1 General Framework for Phoneme Recognition in Continuous Speech

Feature Extraction

This step helps in the process of transforming the input data to features, which defines the properties of various patterns observed in the input. Various feature extraction techniques are available in the literature for extracting the features from the speech input. Some feature extraction methods that gained popularity include Linear Predictive co-efficients (LPC), Poles of Vocal Tract, Reflection co-efficient, Cepstrum, Frequency Cepstral co-efficients, Fast Fourier Transform (FFT), Short-Term Fourier Transform (STFT) and Discrete Wavelet Transform DWT). In LPC, a speech sample at any time is approximated as a sum of previous samples in linear combination. A unique set of coefficients is determined to reduce the sum of squared difference between the predicted values to the actual values. The speech is modelled as a quasi-periodic pulse for voiced signal and random noise for unvoiced speech. They are generally used to model a linear time-varying system.

FFT is an efficient algorithm to evaluate the Discrete Fourier Transform of a signal. It helps to convert a signal from time domain to the frequency domain. It also supports in computing the features like amplitude, power density, phase, magnitude and more. Whereas, STFT is used to compute the sinusoidal frequency and phase information for the speech signal available locally that changes with respect to time. DWT is another feature extraction technique which is a wavelet transform used to capture both the time and frequency domain information of the signal by using discrete set of wavelets with various scales and translations.

Speech Segmentation

This step segments the continuously spoken natural language into phonetic units by recognizing the boundaries of phoneme in the speech. It is a crucial phase which highly affects the accuracy of the model built by influencing the recognition rate. Automatic speech segmentation can be broadly classified as blind segmentation and aided segmentation. Blind segmentation algorithms segment the speech into required units without using any pre-existing, external existing knowledge. They predominantly do rely on acoustic properties of the speech signal and statistical signal analysis. Aided speech segmentation on the other hand uses the historical knowledge of the speech as required for the target speech segments. They are time consuming and few successful methods include the use of Hidden Markov Models (HMM), Dynamic Time Warping (DTW) and Artificial Neural Networks (ANN).

Model Building and Recognition Phase

Once the phonetic units are segmented from the continuous speech, the respective features of all phonetic segments are represented as a feature vectors to build the phonetic dataset. The dataset is then divided into train and test set. The acoustic models are generated with supervised/unsupervised learning techniques using training set. With the acoustic model built, the phonetic segments of any input speech under the context can be classified. HMMs, DTW and ANNs are few acoustic modelling techniques that are well known for its performance in the field of speech recognition.

1.3 REVIEW OF LITERATURE

The research on speech recognition has its root back more than six decades since 1950s [24]. In spite of various methods developed in the literature for the speech recognition problem, Hidden Morkov model based speech recognition had its domination in the field. But with the advent of improvements in technology, in terms of memory, processing power, pattern recognition techniques using artificial intelligence and machine learning, the possibility to

handle and solve complex speech recognition task have increased. This section presents a brief report on various milestones in the area of speech recognition for the past few years.

The author in [25] developed an integrated Support Vector Machine (SVM)/Hidden Markov Model (HMM) for speech recognition that translated the output of the SVM into conditional probabilities and used it as emission probabilities in HMM decoder. The hybrid system was tested against the Defence Advanced Research Project Agency (DARPA) resource management corpus and proved itself promising to the usual Gaussian Mixture Models (GMM). It recorded a word accuracy of 94.10%.

In [26], authors developed a multilingual speech recognition model that had the capability of handling 9 different languages which have minimal overlapping scripts. They have built the union of all grapheme sets of the languages considered and used it in training grapheme based sequence to sequence model that jointly learns the features of all classes in all languages respectively. The method has proved its efficiency by providing 21% improved accuracy when the model is trained without any language related knowledge. An improvement of 7% in accuracy was reported by providing an additional input feature - the language identifier; which highly supported in resolving language clashes.

A speech recognition system robust to noise was developed by the authors in [27] using Hidden Markov Model. The model have used subband coding for subband decomposition and was integrated with the conventional Mel Frequency Cepstral coefficients and subjected to HMM training which yielded attractive classification accuracy for a small closed vocabulary.

AT&T Bell laboratories have developed a CSR system for phonetic transcription [28]. The input speech signal was represented as Cepstral coefficient. The acoustic features were mapped to phonemes using dynamic programming algorithm. The algorithm was implemented as a two-level process where the lower level performs lexical access and the higher level performs the parsing function. DARPA data set was used in building and testing the model. The researchers have obtained an accuracy of 88% for correct word recognition. In some informal tests conducted on speech spoken by the researchers, the word accuracy rate was recorded to about 75%.

Dragon systems developed a 1000 word real time CSR system for a natural large vocabulary problem [29]. The sample rate was considered to be 12 KHz for the input speech signal KHz and filtered at 6 KHz using low pass filter. Speech recognizer was developed as a

HMM, for the mammography recognizing task. 3.4% of word error rate was noticed by the researchers, where as the sentence error rate was reported as 19.5%.

In [30] the authors dealt on the poor performance of held-out test data for a large feed forward neural network. The authors handled the problem of overfitting by randomly omitting half of the feature detectors on each training case. This had prevented complex co-adaptations in which a feature detector was only helpful in the context of several other specific feature detectors. From the collection of complex internal relationships that were available in the given input features, the method identified the most promising features that helped in efficient classification by introducing random dropouts during the learning process. The results show big improvements for many benchmark datasets. The research acts a benchmark for the task of speech recognition and object recognition.

Recent works had shown Deep Neural Networks (DNN) as good choice for practical speech recognition. More detailed study and analysis in [31] showed the strength of DNN in the area of speech recognition. Many factors of DNN in fact influence highly on the accuracy of speech recognition. DNN with a generative pretraining initialized the neural network with a weight space which could then be fine-tuned to achieve a better model. The complexity of training a DNN increased with the number of layers and neurons. This complexity had been reduced by careful initialization of the weight space of the network. The learning process was speeded up by using contrastive divergence. DNN has been viewed as a stack of Restricted Boltzmann Machines (RBMs). The DNN was trained and then it was tied up to HMM which then acted as a classifier. The study highlighted on the performance of DNN on various speech recognition tasks including Bing voice search speech recognition task, switchboard speech recognition task, Google voice input speech recognition task, youtube speech recognition task and English broadcast news speech recognition task. It showed consistency in the performance of DNN-HMMs over GMM-HMMs.

An end to end deep neural network based speech recognition system was developed by the authors in [32] for Tamil speech recognition. It used the state-of-art maximum likelihood linear transformation and speaker-adaptive techniques during its training. It had used 6.5 hours of speech data and analysed the model for phoneme recognition and word recognition. The model achieved 24.9% phone error rate and 3.5% word error rate.

A deep neural network model that performed better phoneme recognition than gaussian mixture model, when applied to TIMIT dataset was proposed in [33]. The deep neural networks

include many layers of features with large number of parameters. The model consists of two phases. In the first phase, the network was pre-trained as multiple layers with window of spectral features without the use of any discriminative information. The second phase involved the use of back-propagation technique that performed discriminative fine tuning.

A novel Context-Dependent (CD) model for Large Vocabulary Speech Recognition (LVSR) was proposed in [34] that implemented Deep Belief Networks (DBN) and context dependent hidden markov model for phoneme recognition. A pre-trained deep neural network hidden markov model is built for the distribution over tied triphone states called senones. The pre-training algorithm was mainly used to initialize deep neural network that helped in optimization and reduction of errors. Business search dataset was used to study the performance of algorithm and showed that it significantly outperformed the conventional context-dependent GMM-HMMs with increase in sentence accuracy.

Many of the resent milestones in speech recognition have shown the strength and power of deep neural networks in achieving comparable and better accuracy to traditional HMMs and GMMs. Deeper networks compared to shallow network modelled the invariable and discriminative features in a better way [35]. The study had visualized DNN as a collection of nonlinear transforms that transformed the input feature into discriminative representations, which was then classified through a log-linear classifier. It had shown that the deeper structure of the network was less sensitive to the smaller input perturbations and resulted in better accuracy.

Investigations on noise robust speech recognition were undergone by the authors in [36]. The researcher introduced a training called noise-aware training. The DNN that was pre-trained with contrastive divergence, used a multi-conditional training to handle the noise variations on speech, and a feature enhancement algorithm based on Cepstral-Minimum Mean Squared Error(C-MMSE). There, the speech recognition on Aurora 4 result showed 7.5% better accuracy when using drop out techniques to DNN learning compared to standard DNN.

Some researchers had also used deep belief networks to pre-train their models. A DBN pre-trained context-dependent ANN/HMM system was modelled and tested on two different datasets in [37]. The performance of the system outperformed the baseline GMM/HMM system in terms of Word Error Rate (WER). The performance of the model was further improved by Maximum Mutual Information (MMI) fine tuning and Segmental Conditional Random Fields (SCARF). The datasets used were large vocabulary continuous speech one with 5780 hours of speech and the other with 1400 hours of speech with 7969 and 17552 target states respectively in

each dataset. The success of Context dependant DNN-HMMs had been also studied in [34]. It discussed the key design choices that performed a major role in the process of recognition. It successfully applied the DNN for pretraining and builds a robust HMM model. The results showed a better phoneme recognition rate to the baseline GMM-HMM.

The nature of DNN forced the adaptation of parameters of a CD-DNN-HMM, a challenging task as discussed in [38] due to the larger number of layers, larger output layer and thousands of neurons in the network. Therefore, Kullback-Leibler divergence regularization was added to the adaptation criterion. The new adaptation technique was applied for various experiments on Xbox voice search, short message dictations, switchboard and lecture speech transcription tasks and showed a reduction in the relative error for speaker independent CD-DNN-HMM systems under both supervised and unsupervised adaptation setups.

Using a traditional speech recognition system required complex learning efforts to capture discriminative features of noisy data, reverberation and speaker variation. But, DNNs had showed up their performance being robust to such data [39]. It did not even require a phonetic dictionary. Huge collection of varied data was obtained using Recurrent Neural Networks (RNN), a set of novel data synthesis techniques and a few GPU. This showed a right direction for speech recognition on languages that were in lack of resources, where many researchers were stuck into. Deep speech system outperformed the commercially available systems by showing a better error rate. It was tested on the full test set of Switchboard Hub5'00.

The author in [40] proposed time delay neural network architecture. The long term temporal dependencies between acoustic events were modelled effectively using recurrent neural networks. But the sequential nature of RNN learning algorithm took higher time for training feed forward networks. The model used sub-sampling method which considerably reduced computation time during training. Experiments were carried out on various LVCSR tasks and result showed the effectiveness of that proposed architecture and proved itself by varying data ranging from 3 to 1800 hours.

Recurrent neural networks were proved to be a powerful model for sequential data [41]. The sequence labelling problems were solved using techniques like connectionist temporal classification for training RNNs. These supported to solve problem even for unknown input-output alignments. The combination of these methods with the long short-term memory RNN architecture had proved its strength in cursive handwriting recognition that delivered state-of-the-art results. The study was also accomplished by combining multi-level representation

capability of deep recurrent neural networks and the capability of RNNs to maintain a longer context range in a flexible manner. It was found that deep Long Short-term Memory RNNs achieved a test set error of 17.7% on the TIMIT phoneme recognition benchmark when it was trained end-to-end with suitable regularisation.

From the literature review, few setbacks that need attention of researchers concerned in Tamil continuous speech recognition were observed. Most of the research on Tamil speech recognition has only focused on digit, word level and minimum vocabulary speech recognition. In spite of few speech corpus available in Linguistic Data Consortium (LDC), acquiring a Tamil speech corpus for research remains a challenge. Once the speech corpus is made available, the next challenge comes with the feature extraction methods. Various feature extraction technique like MFCC, LPC, PLP, Wavelet, RAST, PCA are available where MFCC is found to be dominant. But each has their own pros and cons. It can be observed the wavelet transforms have more pros in their side as it localizes both the time and frequency domain simultaneously and also windowing is dynamic that helps in capturing various levels of signal [42].

Segmentation of continuous speech is another challenge. The complexity of segmentation is affected with respect to various characteristics of speech like articulation, pronunciation, disfluency, pitch, pauses, speech rate, etc. Various supervised segmentation methods are available but still face the problem of reference corpus. Availability of labelled word level, syllable level of phoneme level dataset for the language considered is again a challenge. Building such reference dataset manually with experts is a time consuming and expensive task. Whereas, unsupervised segmentation algorithms are need to be explored more. The accuracy of the unsupervised segmentation algorithms is evident for achieving efficient recognition models. It can also be observed that building a phoneme recognition model highly relies on the spoken language. Thus, requires developing language specific models for achieving more recognition accuracies. Eventhough, various phoneme recognition models have been built using HMM, GMM, ANN, CNN, RNN, etc., only limited focus has been given to build phoneme recognition model for Tamil continuous speech using deep learning. With these limitations, the research proposes to take up few challenges and address those problems. A summary of the literature review is presented below.

Table I Summary of Literature Survey

Author and Year	Dataset	Objective	Algorithm or Methodology	Results
Krüger, S. E., Schafföner, M., Katz, M., Andelic, E., & Wendemuth, A. (2005)	DARPA resource management corpus	Word recognition	Integrated SVM/HMM	Reported 94.10% accuracy
Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., & Rao, K. (2018)	Utterances of 9 Indian languages collected through desktop and mobile devices	Grapheme based recognition	Sequence-to-sequence multilingual model using LSTM cells	21% improvement in accuracy of models built for corresponding languages
S. E. Leninson, A. Ljolie, L.G. Miller, (1990)	DARPA resource management task	Phoneme recognition and word recognition	HMM with dynamic programming	88% word accuracy
Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012)	TIMIT	Phoneme recognition	HMM with neural network model with dropout	42.4% recognition rate
Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B. and Sainath, T., (2012)	Switchboard, English broadcast news, Bing voice search, Google voice input, Youtube	Phoneme recognition	DNN-HMM models	DNN-HMM outperform respective GMM-HMM
Mohamed, Abdel-rahman, George E. Dahl, and Geoffrey Hinton (2011)	TIMIT	Phoneme recognition	DBN	20.7% Phoneme error rate
Yu, D., Seltzer, M. L., Li, J., Huang, J. T., & Seide, F. (2013)	Switchboard	Word recognition with senones	DNN	Invariant and discriminative features are extracted better with DNNs

Seltzer, Michael L., Dong Yu, and Yongqiang Wang (2013)	Aurora 4 based on Wall Street journal corpus	Noise robust acoustic model	Noise-aware DNN	12.4% WER
Jaitly, N., Nguyen, P., Senior, A., & Vanhoucke, V. (2012)	Voice search system, Youtube	Acoustic model	DBN pretrained ANN/HMM + Maxmimum mutual information + Segmental conditional random fields	11.8% Word Error Rate (WER) for voice search system and 46.2% word error rate for Youtube
Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012)	Business search dataset from Bing mobile voice search application	Phoneme recognition	CD-DNN-HMM	69.5% accuracy and outperformed CD-GMM-HMM models
Yu, D., Yao, K., Su, H., Li, G., & Seide, F. (2013)	Xbox voice search, short message dictations, switchboard, lecture speech	Senone model	Kullback-Leibler divergence regularized DNN	20.2% WER
Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A. and Ng, A.Y., (2014)	Wall street journal, switchboard, fisher, baidu	End-to-end speech recognition	Deep Speech using RNN	11.8% WER
Peddinti, Vijayaditya, Daniel Povey, and Sanjeev Khudanpur (2015)	Switchboard	Modeling long-term temporal dependencies for speech recognition	Time-delay neural network	14% WER
Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. (2013)	TIMIT	Phoneme recognition	Deep RNN	17.7% Phoneme Error Rate (PER)

Motivation

Speech recognition is catered to a wide range of application areas. Users of various domains will be much benefitted if these applications are available in their regional language. Eventhough the speech recognition has its footsteps back since six decades, proper benchmark datasets are not currently available for undergoing speech research in most of the South Indian Languages. Speech data is observed to be highly variable due to the co-articulating effects of the neighbouring phonemes in continuous speech and by the speech characteristics of various speakers. There is a need for automatic segmentation and labelling of Tamil continuous speech to enable this research community in building labelled phonetic datasets. Also, acoustic models need to be built that is capable of capturing the high variability in the speech data to perform phoneme recognition of Tamil continuous speech. Here the problem to build accurate acoustic models for the ancient Tamil language is considered to improve the consumption of the advantages gained through speech recognition by the huge Tamil speaking community living around the globe. To promote the continuous speech recognition for Tamil language, as a first and foremost effort, a speech corpus needs to be developed. Next, an unsupervised segmentation method that helps to segment the continuous speech to the required level irrespective of speaker variability. The robust segmentation algorithm has to be tailored for building the dataset necessary to build efficient acoustic model. In spite of several existing models for phoneme recognition, there is a gap observed in the area of Tamil continuous speech recognition that need to be fulfilled with the contemporary powerful techniques in deep learning. Thus, this research proposes to build precise acoustic models for phoneme recognition in Tamil continuous speech using deep learning.

1.4 OBJECTIVES OF THE RESEARCH

The main objective of this research is to develop a framework using deep learning approaches to efficiently recognize the phonemes of Tamil continuous speech. The overall objective is split and listed out as a line of investigation objectives to achieve the research objective.

- To build a speech corpus for Tamil language that includes the speech signals, its transcripts and phoneme database.
- To develop an efficient algorithm to automatically segment the Tamil continuous speech into phonetic segments using graph cut based segmentation algorithm.

- To design and develop a methodology to build an acoustic model for Tamil phoneme recognition performing with good classification accuracy using deep belief networks.
- To design a framework that enables to build the acoustic model with less time complexity and to optimize the acoustic model for better phoneme classification using Particle Swarm Optimization(PSO) pre-trained deep belief networks.
- To enhance the efficiency of the acoustic model by considering the data imbalance problem while building the acoustic model with Weighted Mean Square Error measure.

The phoneme classification problem is formulated as pattern recognition task and solved using unsupervised segmentation and modern deep learning technique. A new Tamil speech corpus called Kazhangiyam is developed and the corresponding dataset is created using DWT features. Phoneme recognition models are built using deep belief networks with various pretraining phases for better parameter optimization and with the proposed weighted mean square error loss function.

1.5 ORGANIZATION OF THESIS

The thesis is organized into nine chapters. The rest of the thesis is structured as follows.

- Chapter 2 gives a background on machine learning, deep learning architectures and methods that supported this research in building the phoneme recognition models.
- Chapter 3 explains the proposed model for continuous speech recognition, creation of Tamil speech corpus – 'Kazhangiyam'. The preprocessing tasks like filtering, feature extraction, segmentation and dataset preparation are also elucidated.
- Chapter 4 elucidates on the pilot study undergone with Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference System Tamil phoneme recognition. The experiments conducted and the results obtained are discussed.
- Chapter 5 describes a phoneme recognition framework that uses graph cut based segmentation for segmenting the continuous speech into phonemes and deep belief network for building Tamil phoneme recognition model. The results for the experiments are analyzed and presented.
- Chapter 6 discusses the Particle swarm optimization pre-trained Deep Belief Networks Tamil phoneme recognition models. It also elucidates the respective experiments and their comparative analysis with contrastive divergence based DBN acoustic model.

- Temperature controlled PSO, a modified version of PSO is introduced in Chapter 7. Procedure to build TPSO-DBN acoustic model is explained. It also discusses the experiments on TPSO-DBN acoustic model and presents a comparative analysis.
- In Chapter 8, a loss function for imbalanced multiclass dataset termed as Weighted Mean Square Error (WMSE) is proposed. This chapter discusses the need for a specific function capable of handling the imbalanced dataset and defines the proposed WMSE. The WMSE loss function is validated in pre-training and training phases of DBN model building framework.
- Chapter 9 summarizes the findings of the research and the contributions made in building efficient Tamil phoneme recognition models. It also presents the scope for further research.