Chapter 1
Introduction
1.1 OVERVIEW OF SPEECH RECOGNITION SYSTEM

Man-machine communication using speech leaves hands and eyes free to perform other tasks. For the most common tasks, speech is a much more natural way to communicate than typing on a keyboard. Majority of people learn to speak long before they learn to write or type. More than three percent of all Americans and Germans cannot read or write at all (World Factbook, 1996) and a situation in India is worst. Since word processing is the most common use of PCs, the potential market for fast dictation system is large. Many companies such as Microsoft (Alleva, 1996), IBM (Gopalkrishnan, 1995), Dragon (Ellermann, 1993) and Siemens (Niemoller, 1997) put substantial resources into the development of such products. Other than being more economical, computers can also be programmed to be more discreet than humans. Applications that require this sort of discretion could be the dictation of business letters, speech-to-speech translation (Osterholtz, 1992), (Bub, 1997) and monitoring of telephone lines. To exploit the sources like tapes, on which the broadcast news, interviews and talk shows are stored, a summary or a keyword index is required to search speech documents, the same way as collections of filed text documents. Using speech recognition system, such databases can be created and maintained automatically (Hauptmann, 1997). The space requirements of a microphone are substantially lower than those of a full size keyboard. Considering that keyboards tend to be among the largest parts of a modern computer this could also become an issue when designing the next generation of palm-tops.

In general, a speech recognizer consists of three parts; feature analysis, unit matching system and postprocessor (Pan, 1985) (Rabiner, 1989) (Fallside, 1992). A schematic representation is depicted in Figure 1.1.

Feature Analysis: In this part of a speech recognition system, the incoming speech signal is analyzed and transformed into a set of vectors containing the results of the spectral and/or temporal analysis of the signal. Different features which are normally extracted from the speech signal are linear predictive coefficients (LPC) (Makhoul, 1975) (Rabiner, 1978a), linear predictive cepstrum

![Figure 1.1: A speech recognition system.](image)

**Unit Matching System:** Units which are to be recognized by the unit matching system can vary from whole word units to sub-word units (syllables) or even to smaller components of speech like phonemes. Generally, a more complex unit, such as a word, is easier to recognize, whereas the advantage of a less complex unit is that there are fewer of such units in a language. Unit recognition can be implemented using deterministic classifiers (minimum distance, discriminative, neural networks, etc.) (Huang, 1990) (Gold, 2000). Another approach that captures the dynamic structure of speech is dynamic classifiers (dynamic time warping (DTW), hidden Markov models (HMM), time-delay neural networks (TDNN) or hybrid recognition systems which combines neural networks with DTW or HMM).

**Postprocessor:** The postprocessing part in speech recognition system deals with the units recognized by unit matching system. The knowledge about the acoustical structure of the words (which are to be recognized), as well as about the grammatical structure of the speech is used to take the final decision.

### 1.2 A REVIEW OF SPEECH RECOGNITION RESEARCH

The earliest attempts to devise systems for automatic speech recognition by machine were made in the 1950s, when various researchers tried to exploit the fundamental ideas of acoustic-phonetics. In 1952, at Bell Laboratories, Davis,
Biddulph and Balashek built a system for isolated digit recognition for a single speaker (Davis, 1952). The system relied heavily on measuring spectral resonances during the vowel region of each digit. In an independent effort at RCA Laboratories in 1956, Olson and Belar tried to recognize 10 distinct syllables of a single talker, as embodied in 10 monosyllabic words (Olson, 1956). The system again relied on spectral measurements (as provided by analog filter bank) primarily during vowel region. In 1959, at University College in England, Fry tried to build a phoneme recognizer to recognize four vowels and nine consonants (Fry, 1959). He used a spectrum analyzer and a pattern matcher to make the recognition decision. A novel aspect of this research was the use of statistical information about allowable sequences of phonemes in English to improve overall phoneme accuracy for words consisting of two or more phonemes. Another effort of note in this period was the vowel recognizer of Forgie and Forgie, constructed at MIT Lincoln Laboratories in 1959, in which 10 vowels were recognized in a speaker independent manner (Forgie, 1959). Again a filter bank analyzer was used to provide spectral information, and a time varying estimate of the vocal tract resonances was made to decide which vowel was spoken.

In the 1960s, several fundamental ideas in speech recognition surfaced and were published. However, the decade started with several Japanese laboratories entering the recognition arena and building special purpose hardware as part of the Radio Research Lab in Tokyo (Suzuki, 1961), and that was hardware vowel recognizer. Another hardware effort in Japan was the work of Sakai and Doshita of Kyoto University in 1962, who built a hardware phoneme recognizer (Sakai, 1962). A hardware speech segmenter was used along with zero-crossing analysis of different parts of the spoken input to provide the recognition output. A third Japanese effort was the digit recognizer hardware of Nagata and coworkers at NEC Laboratories in 1963 (Nagata, 1963). In the 1960s, three key research projects were initiated that has had major implications on the research and development of speech recognition for the past forty years. The first of these projects was the efforts of Martin and his colleagues at RCA Laboratories, beginning in the late 1960s, to develop realistic solutions to the problems
associated with non-uniformity of time scales in speech events. Martin developed a set of elementary time normalization methods based on the ability to reliably detect speech starts and ends that significantly reduced the variability of the recognition scores (Martin, 1964). Martin ultimately developed the method and founded one of the first companies. At about the same time, in Soviet Union, Vintsyuk proposed the use of dynamic programming methods for time aligning a pair of speech utterances (Vintsyuk, 1968). Although the essence of the concepts of dynamic time warping, as well as rudimentary versions of the algorithms for connected word recognition, were embodied in Vintsyuk’s work, it was largely unknown in the West and did not come to light until the early 1980s; this was long after the more formal methods were proposed and implemented by others. A final achievement of note in the 1960s was the pioneering research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes (Reddy, 1966).

In the 1970s speech recognition research achieved a number of significant milestones. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies by Velichko and Zagoruyko in Russia (Velichko, 1970), Sakoe and Chiba in Japan (Sakoe, 1978), and Itakura in the USA (Itakura, 1975). The Russian studies helped to advance the use of pattern recognition ideas in speech recognition; the Japanese research showed how dynamic time warping (DTW) methods could be successfully applied; and Itakura’s research showed how the ideas of linear predictive coding (LPC), which had already been successfully used in low-bit-rate speech coding, could be extended to speech-recognition systems through the use of an appropriate distance measure based on LPC spectral parameters.

Another milestone of the 1970s was the beginning of a longstanding, highly successful group effort in large vocabulary speech recognition at IBM in which researchers studied three distinct tasks over a period of almost two decades, namely the New Raleigh language (Tappert, 1971) for simple database queries, the laser patent text language (Jelinek, 1975) for transcribing laser patents, and the
office correspondence task, called Tangora (Jelinek, 1985), for dictation of simple memos.

In AT&T Bell Labs, researchers began a series of experiments aimed at making speech recognition system that were truly speaker independent (Rabiner, 1979). To achieve this goal a wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population. This research has been refined over a decade so that the techniques for creating speaker independent patterns are well understood and widely used.

Just as isolated word recognition was a key focus of research in the 1970s, the problem of connected word recognition was a focus of research in the 1980s. Here the goal was to create a robust system capable of recognizing a fluently spoken string of words (e.g., digits) based on matching a concatenated pattern of individual words. A wide variety of connected word recognition algorithms were formulated and implemented, including the two-level dynamic programming approach of Sakoe at Nippon Electric Corporation (NEC) (Sakoe, 1979), the one-pass method of Bridle and Brown at Joint Speech Research Unit (JSRU) in England (Bridle, 1979), the level building approach of Myers and Rabiner at Bell Labs (Myers, 1981), and the frame synchronous level building approach of Lee and Rabiner at Bell Labs (Lee, 1989). Each of the optimal matching procedure had its own implementational advantages, which were exploited for a wide range of tasks.

Speech research in the 1980s was characterized by a shift in technology from template based approaches to statistical modeling methods especially the hidden Markov model approach (Ferguson, 1980) (Rabiner, 1989). Another technology that was reintroduced in the late 1980s was the idea of applying neural networks to problems in speech recognition. Neural networks were first introduced in the 1950s, but they did not prove useful initially because they had many practical problems. In the 1980s, however, a deeper understanding of the strengths and
limitations of the technology was obtained, as well as the relationships of the
technology to classical signal classification methods were established. Several
new ways of implementing neural network systems were also proposed

The 1980s was a decade in which a major impetus was given to large vocabulary,
continuous speech recognition systems by the Defense Advanced Research
Projects Agency (DARPA) community. Major research contributions resulted
from efforts at CMU notably the SPHINX system (Lee, 1990), BBN with the
BYBLOS system (Chow, 1987), Lincoln Labs (Paul, 1989), SRI (Weintraub,
1989), MIT (Zue, 1989) and AT&T Bell Labs (Lee, 1990). The DARPA program
has been continued into the 1990s, with emphasis shifting to natural language
front ends to recognizer and the task shifting to retrieval of air travel information.
At the same time, speech technology has been increasingly used within telephone
networks to automate as well as enhance operator services. Similarly, DARPA
program moved on to a task referred to as Broadcast News, in that the automatic
transcription of broadcast data is closely related to several potential commercial
applications.

In the past decade, there was significant work to improve recognition robustness
to different channels, as well as to acoustic noise (Gales, 1995) (Stern, 1996) (Lee,
1999) (Gold, 2000). Hermansky and Morgan have reported RASTA (relative
spectra) processing technique with discussion on relationship with human
auditory perception, effect of additive noise and convolution noise (Hermansky,
1994). There has been an increased emphasis on issues of pronunciation (Riley,
1996), dialog modeling (Cohen, 1997), long distance dependencies within word
sequences (Rosenfeld, 1996) and acoustic model adaption (Leggetter, 1995). A
rapid expansion of research related to speech recognition has been done by
developing methods and systems for speaker identification and verification
(Reynolds, 1995) (Furui, 1996), as well as for language identification
(Muthusamy, 1994). Ramchandran, Zilovic and Mammone have reported various
linear predictive analysis methods and compared from the point of view of
robustness to noise and of application to speaker identification (Ramchandran, 1995). Rajendran and Yegnanarayana have proposed an algorithm based on F0 patterns to hypothesize word boundaries and function words in continuous speech in Hindi (National language of India). They have also reported the robustness of the algorithm in handling noisy speech input conditions and telephone speech (Rajendran, 1996). Umesh, Cohen, Marinovic and Nelson have used scale transform of spectral envelop of speech utterances by different speakers for reducing variations in acoustic features (Umesh, 1999). For vowel classification and isolated digit recognition, Umesh et al. have used mel cepstrum based features and scale transform based features and reported that scale transform based features provide the better recognition performance under noisy condition (Umesh, 1999a).

With the advent of backpropagation, a training technique for multilayer perceptrons, in the early 1980s, the neural network field experienced resurgence. Neural network with backpropagation learning algorithm learns by changing its connection weights without extra memory overhead. But the major drawback of the backpropagation model is that, it cannot assure a successful learning due to the local minimum problem. Besides, the training time of backpropagation model is frequently too long to stand. One application of neural networks to speech recognition in that period was the use of a committee machine to judge whether a section of speech was voiced or unvoiced (Gevins, 1984). Time delay neural network (TDNN) has been used first time by Makino for the recognition of consonants (Makino, 1983). Waibel et al. have expanded the technique of TDNN by adding the delayed versions at multiple layers in neural network (Waibel, 1989). In mid-1980s, other researchers used Hopfield network for speech recognition (Lippmann, 1987a). Nowlan has used radial basis function network for classification of vowels (Nowlan, 1990). Some researchers have proposed the hybrid speech recognition techniques which contain DTW or HMM to search possible space of word strings and neural networks as a phonetic probability estimator (Morgan, 1995) (Robinson, 1996). Application of different neural networks with different approaches can be found in the work published by

1.3 RELATED LITERATURE SERVEY
Itakura's research showed how the ideas of linear predictive coding (LPC) could be extended to speech recognition systems through the use of log likelihood distance measure based on LPC spectral parameters (Itakura, 1975). He proposed an algorithm to find the best match between the input pattern and reference pattern using dynamic programming technique. Sakoe and Chiba have reported an optimum dynamic programming based time normalization algorithm for spoken word recognition (Sakoe, 1978). They have proposed symmetric form and asymmetric form of algorithms and showed that symmetric form of algorithm exhibits better results. Rabiner et al. (Rabiner, 1978) have proposed DTW algorithms, which relaxed the time registration constraints at the endpoints of test and reference utterances and allowed the dynamic path to follow a locally optimum path at each frame. Myers et al. (Myers, 1980) proposed an approach to DTW for isolated words in which both the reference and test patterns are linearly warped to a equal length and a simplified DTW algorithm is used to handle the nonlinear component of the time alignment.

A. Waibel and B. Yegnanarayana (Waibel, 1983) have reported that the performance of DTW algorithm or nonlinear time alignment algorithm is vocabulary dependent by investigating the performance of different algorithms for two different vocabularies, namely, digits (zero, one, ..., nine) and a highly confusable subset of the alphabet (b, d, e, g, p, v, z). Sadaoki Furui has proposed the staggered DTW algorithm for speaker independent isolated Japanese city names recognition and reported that his algorithm needs comparatively less number of computations for overall distance computations (Furui, 1986). Silverman and Morgan have published tutorial on the use of DTW algorithms for speech recognition (Silverman, 1990). J-C Junqua has used DTW algorithm in SmarTspell™ system for name retrieval over the telephone (Junqua, 1997).
For confusing words (similar sounding words) recognition, Tribolet et al. have used two-pass discrimination model with long time features and short time features (Tribolet, 1982). They have used 39-word vocabulary of the 26-English alphabets, 10-digits and 3-command words and reported that improved feature analysis shows a small improvement in recognition accuracy (0-1.3 percent), whereas the two-pass decision rule provides somewhat better improvement (1-4.4 percent) over traditional LPC based speech recognition model. Rabiner and Wilpon have proposed weighting function based on training data (Rabiner, 1987) (Rabiner, 1993), derived on the basis of Fisher's linear discriminative analysis (Duda, 1973) for the recognition of confusing spoken words. A number of speech recognition techniques using DTW, HMM and neural networks have been developed to deal with confusing words either in isolation or in hybrid form (Euler, 1990) (Cole, 1991) (Junqua, 1991) (Anglade, 1993) (Schmid, 1993).

Various researchers have used predictive neural networks for recognition of spoken words (Iso, 1990) (Iso, 1991) (Tebelskis, 1990) (Tebelskis, 1991) (Mellouk, 1993). In training of predictive neural network, the error is backpropagated hence the time required to train the neural network is more with associated drawbacks of backpropagation training algorithms. In Tebelskis predictive neural networks, two major weaknesses were: a lack of discrimination and inconsistency between training and testing criteria.

Tan Lee et al. have proposed modular recurrent neural networks for isolated word recognition (Lee, 1998). They have used 20-English words, 11- Cantonese digits and 58-Cantonese syllables as vocabulary. To recognize an input word, the best matching word is determined based on its temporal output response. They have used separate recurrent neural network for each word in the vocabulary. Chandra Sekhar and Siva Rama Krishna Rao have used multiple modular networks for the recognition of utterances of consonant-vowel units of Indian languages (Chandra Sekhar, 1999). They have trained separate neural networks (subnets) for subsets of classes and the outputs of the subnets are processed to perform the classification.

For Chinese spoken word recognition, researchers in Institute of Information Engineering, National Cheng Kung University, Tainan, Taiwan have used the fuzzy neural networks and reported that fuzzy based neural networks show stronger reliability on classification with respect to probabilistic neural networks and backpropagation (Kao, 1992) (Kuo, 1994). Grigore and Gavat used fuzzy neural network for the recognition of Romanian vowels (Grigore, 1999). They have used a fuzzy neural network model based on the classical multilayer perceptron and a fuzzy classifier based on Kohonen's self organizing neural network. And reported that both the models show better recognition results because of introduction of fuzzy concept in classifiers.

Nikola Kasabov and Michael Watts of University of Otago, New Zealand, have used fuzzy neural network with online adaptive learning for phoneme recognition in the system they called EFuNNs (Kasabov, 1999). The system accommodates new input data including new features, new classes, etc. Both feature based similarities and temporal dependencies that are present in the input data (new pronunciation, new accent) are learned and stored in the connections and adjusted over time. Kasabov have used the unsupervised learning algorithm for clustering of spoken words (Kasabov, 2001).
1.4 PROBLEM DEFINITION
The objective of this thesis is to study some of the existing algorithms of deterministic classifiers for speech recognition and to propose new algorithms. In particular, this thesis addresses to the recognition of isolated spoken Marathi (Language spoken in the state of Maharashtra, India) words using minimum distance classifiers, discriminative weighting classifiers and neural network classifiers.

1.5 ORGANIZATION OF THE THESIS
The thesis is organized as follows.

Chapter 1 discusses need of speech recognition with overview of speech recognition system, review of speech recognition research, related literature survey and problem definition.

Chapter 2 describes a nonlinear time alignment algorithm (TAA) and an alternative approach (DW_PRO) to the discriminative weighting method proposed by Rabiner and Wilpon (DW_R) (Rabiner, 1987) (Rabiner, 1993), based on Fisher’s linear discriminative analysis (Duda, 1973) for the recognition of confusing spoken words. The vocabulary used in speaker dependent mode is 46-isolated monosyllabic phonetically confusing Marathi alphabets and in speaker independent mode is 10-Marathi digits //shunya// (0) to //nau// (9). The performance of TAA to speaker dependent mode (Doye, 2001) and independent mode (Doye, 2001a) is compared with well-established DTW algorithms and found that TAA shows better accuracy, needs less number of computations and shows robustness to endpoint detection and noise. The reasons for better performance of TAA are discussed in this chapter.

The performance of DW_PRO (Doye, 2001b) is tested with various speech parameters such as LPC, MFCC, LFCC and regression coefficients of MFCC and LFCC and found that DW_PRO shows better performance than DW_R. The reasons for better performance of DW_PRO are discussed in this chapter.
In chapter 3, we propose the modular fuzzy min-max neural network (MMFMM). MMFMM contains number of modules of modified fuzzy min-max neural network (MFMM) (Doye, 2002). In MFMM, the transfer function of output layer of general fuzzy min-max neural network (GFMM) (Gabrys, 2000) is modified. Similarly, learning algorithm is also modified. The performance of GFMM, MFMM and MMFMM to recognition of spoken Marathi digits is presented. In MMFMM (Doye, 2002a), words are segmented into $f$ segments and individual segment is used for training the $f$ modules. In the testing of input word, the average of $d$ best fuzzy membership values of hyperboxes of all classes for each segment is computed. Then average of $f$ fuzzy membership values of each class is computed. The class showing largest fuzzy membership value is assigned to input word. In training as well as in testing, every segment is treated as separate pattern and every segment contribute to classification decision, which is essential in multi-syllabic words. MMFMM has shown better average recognition accuracy than GFMM.

Chapter 4 describes a modified fuzzy hypersphere neural network (MFHS) (Doye, 2002b) with two learning algorithms for the recognition of isolated Marathi digits. MFHS is a modification of fuzzy hypersphere neural network (FHS) proposed by Kulkarni and Sontakke (kulkarni, 2001b). The architecture is modified which uses only three layers instead of four. The transfer function of the output layer neurons is also modified. In the learning algorithm of FHS, the first pattern contained by the hypersphere is the center of hypersphere and the hypersphere grows around the center with training patterns associated with the hypersphere. Hence, center of hypersphere is not a centroid of the patterns associated with the hypersphere. The membership of the testing pattern with the hypersphere is a function of center of hypersphere; therefore the recognition rate is less, which improves in improved learning algorithm called as two-stage learning algorithm. The hyperspheres are created in the first stage. The centroids of the patterns associated with the hyperspheres are computed by using the mean of the features. The computed centroids may not be the real training patterns. In the second stage, the hyperspheres are initialized with the computed centroids and all the training
patterns are applied for training. The hyperspheres, which are points in the hyperspace, are not used for the initialization as they may be accommodated by the hyperspheres in the second stage.

In another learning algorithm, called as centroid-based learning algorithm, the hyperspheres are created and are modified during learning such that the center of hypersphere will be a centroid of the patterns associated with the particular hypersphere continuously. These algorithms exhibit better performance than FHS.

Chapter 5 describes general fuzzy hypersphere neural network (GFHS) (Kulkarni, 2002) that uses supervised and unsupervised learning within a single training algorithm. It is an extension of FHS (Kulkarni, 2001b) and can be used for pure classification, pure clustering or hybrid classification/clustering. The performance of GFHS is compared with GFMM for classification and unsupervised clustering of patterns of spoken Marathi digits and found that GFHS exhibits comparable response.

Finally, chapter 6 gives conclusions and direction for future work.