Automatic Speech Recognition

Samudravijaya K

Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005
Email: chief@tifr.res.in

1 Introduction

Information processing machines have become ubiquitous. However, the current modes of human machine communication are geared more towards living with the limitations of computer input/output devices rather than the convenience of humans. Speech is the primary mode of communication among human beings. On the other hand, prevalent means of input to computers is through a keyboard or a mouse. It would be nice if computers could **listen** to human speech and carry out their commands. Automatic Speech Recognition (ASR) is the process of deriving the transcription (word sequence) of an utterance, given the speech waveform. Speech understanding goes one step further, and gleans the meaning of the utterance in order to carry out the speaker's command. This article gives a tutorial introduction to ASR. Reader may refer to [1] for an overview of speech recognition and understanding.

This article is organised as follows. This section mentions salient application areas of ASR and lists the types of speech recognition systems. After describing basic steps of production of speech sounds, Section 2 illustrates various sources of variability of speech signal that makes the task of speech recognition hard. Section 3 describes the signal processing, modeling of acoustic and linguistic knowledge, and matching of test pattern with trained models. A small sample of ASR application systems in use in India and abroad is given in Section 4. Section 5 lists some serious limitations of current ASR models, and briefly discusses the challenges ahead on the road to realisation of natural and ubiquitous speech I/O interfaces. It also mentions a few Indian efforts in this direction. Section 6 draws some conclusions.

1.1 Applications areas of ASR

ASR systems facilitate a physically handicapped person to command and control a machine. Even ordinary persons would prefer a voice interface over a keyboard or mouse. The advantage is more obvious in case of small hand held devices. Dictation machine is a well known application of ASR. Thanks to the ubiquitous telecommunication systems, speech interface is very convenient for data entry, access of information from remote databases, interactive services such as ticket reservation. ASR systems are expedient in cases where hands and eyes are busy such as driving or surgery. They are useful for teaching phonetic and programmed teaching as well.

1.2 Types of ASR

Speech Recognition Systems can be categorised into different groups depending on the constraints imposed on the nature of the input speech.

- **Number of speakers**: A system is said to be *speaker independent* if it can recognise speech of any and every speaker; such a system has learnt the characteristics of a large number of speakers. A large amount of a user's speech data is necessary for training a *speaker dependent* system. Such a system does not recognise other's speech well. *Speaker adaptive* systems, on the other hand, are speaker independent systems to start with, but have the capability to adapt to the voice of a new speaker provided sufficient amount of his/her speech is provided for training the system. Popular dictation machine is a speaker adapted system.

- **Nature of the utterance**: A user is required to utter words with clear pause between words in an *Isolated Word Recognition* system. A *Connected Word Recognition* system can recognise words, drawn from a small set, spoken without need for a pause between words. On the other hand, *Continuous Speech Recognition* systems recognise sentences spoken continuously. *Spontaneous speech* recognition system can handle speech disfluencies such as ah, am or false starts, grammatical errors present in a conversational speech. A *Keyword Spotting System* keeps looking for a pre-specified set of words and detects the presence of any one of them in the input speech.

- **Vocabulary size**: An ASR system that can recognise a small number of words (say, 10 digits) is called a small vocabulary system. Medium vocabulary systems can recognise a few hundreds of words. Large and Very Large ASR systems are trained with several thousands and several tens of thousands of words respectively. Examples application domains of small, medium and very large vocabulary systems are telephone/credit card number recognition, command and control, dictation systems respectively.

- **Spectral bandwidth**: The bandwidth of telephone/mobile channel is limited to 300-3400Hz and therefore attenuates frequency components outside this passband. Such a speech is called *narrow-band* speech. In contrast, normal speech that does not go through such a channel is call *wideband* speech; it contains a a wider spectrum limited only by the sampling frequency. As a result, recognition accuracy of ASR systems trained with wideband speech is better. Moreover, an ASR system trained with narrow band speech performs poorly with wideband speech and vice versa.

2 Why speech recognition is difficult?

Despite decades of research in the area, performance of ASR systems is nowhere near human capabilities. Why Speech Recognition is so difficult? This is primarily due to variability of speech signal.

Speech Recognition is essentially a decoding process. Figure 2 illustrates the encoding of a message into speech waveform and the decoding of the message by a recognition system. Speech can be modelled as a sequence of linguistic units called phonemes. For example, each character of Devanagari alphabet essentially represents a phoneme. In order to better appreciate the difficulties associated with ASR, it is necessary to understand the production of speech sounds and sources of variabilities.

2.1 Production of speech sounds

A knowledge of generation of various speech sounds will help us to understand spectral and temporal properties of speech sounds. This, in turn, will enable us to characterise sounds in terms of features which will aid in recognition and classification of speech sounds.

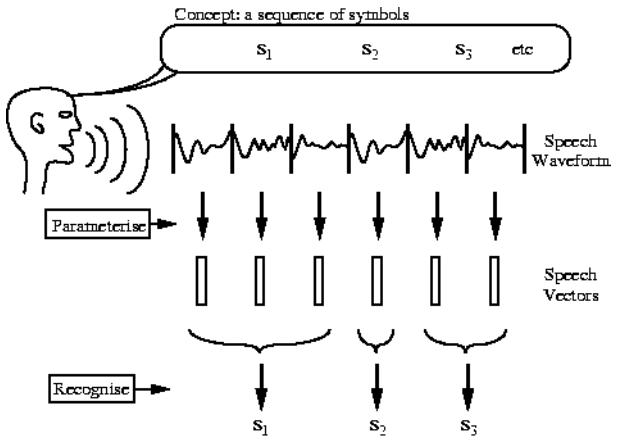

Figure 1: Message encoding and decoding. (Source: [2])

Sounds are generated when air from the lungs excite the air cavity of the mouth. Figure 2 shows the human anatomy relevant to speech production. In case of production of a voiced sound, say vowel /a/, the glottis opens and closes periodically. Consequently, puffs of air from lungs excite the oral cavity. During the closure periods of the glottis, resonances are set up in the oral cavity. The waveform coming out of the lips has the signature of both the excitation and the resonant cavity. The frequency of vibration of the glottis is popularly known as the pitch frequency.

For the production a nasal sound, the oral passage is blocked, and the velum that normally blocks the nasal passage is lifted. During the production of unvoiced sounds, the glottis does not vibrate and is open. The oral cavity is excited by aperiodic source. For example, in the production of /s/, air rushing out of a narrow constriction between the tongue and upper teeth excites the cavity in front of the teeth.

In order to produce different sounds, a speaker changes the size and shape of oral cavity by movement of articulators such as tongue, jaw, lips. The resonant oral tract is generally modelled as a time-varying linear filter. Such a model of speech production is called a **source-filter model**. The excitation source can be periodic (as in case of voiced sounds) or aperiodic (example: /s/) or both (example: /z/).

For the vowel /a/, the vocal tract can be approximated, during the closed phase of glottis vibration, as a uniform tube closed at one end. The fundamental mode of resonance corresponds to a quarter wave. If we assume 340m/s as the speed of sound in air and 17 cm as the length, L of the vocal tract from glottis to lips, the fundamental frequency of resonance can be calculated as

$$\nu = c/\lambda = c/(4 * L) = 34000/4 * 17 = 500Hz \tag{1}$$

The frequencies of odd harmonics are 1500Hz, 2500Hz etc. The spectrum of glottal vibration is a line spectrum with peaks at 100, 200, 300Hz etc. if the pitch frequency is 100 Hz. From the theory of digital filters, it can be easily shown that the log power spectrum of the output of the filter (speech wave) is the sum of the log spectra of source and filter. Figure 3 shows these 3 spectra for neutral vowel /a/, for the ideal case as represented by Eqn. 1. Top and bottom panels of the figure correspond to cases when the pitch ($F0$) is 100 and 200Hz respectively. Notice that although the spectrum of the speech waveforms (right figures) appear slightly differently due to different pitch, both correspond to the same vowel. Thus, variation in speech spectrum due to different pitch should be ignored while doing speech recognition.

Figure 4 shows actual power spectra of two speech sounds of the Hindi word “ki” on a log scale. The

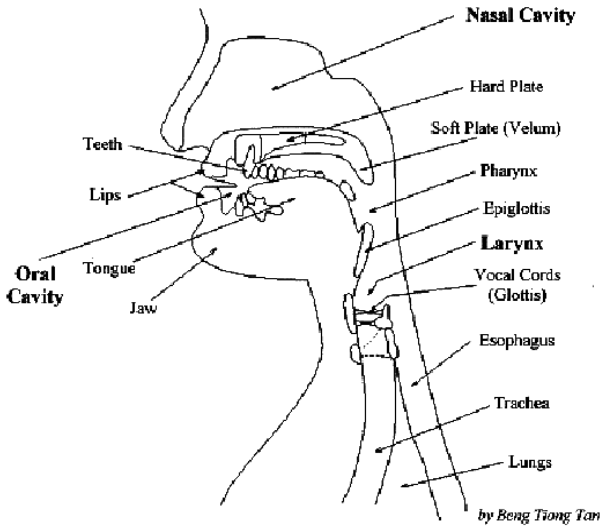

Figure 2: Human vocal system (Source: [3])

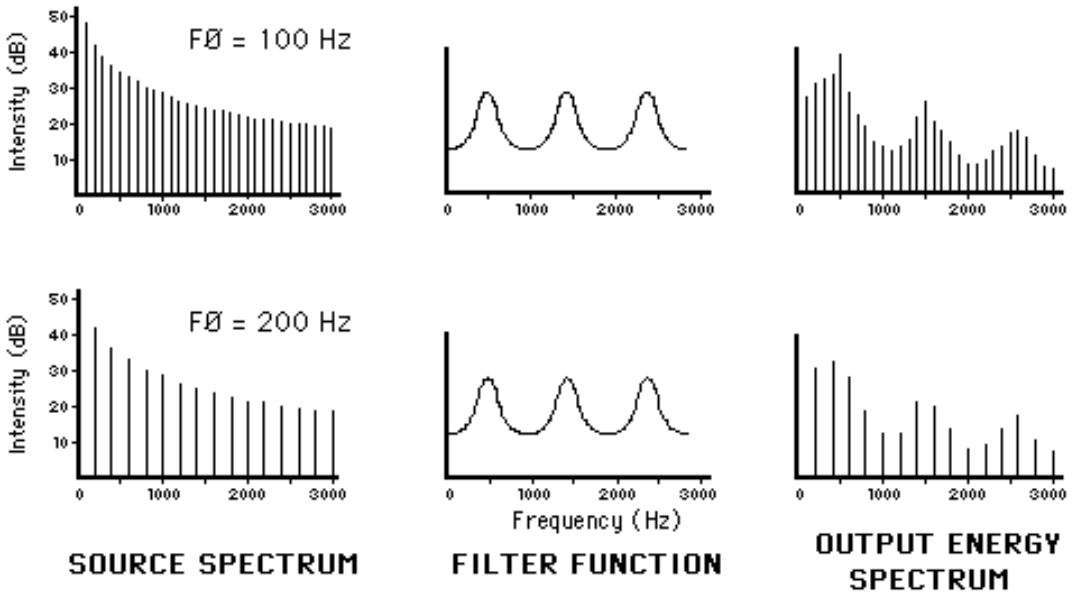

Figure 3: Ideal spectra of source, filter and output speech for vowel /a/. Top and bottom panels correspond to cases when the pitch is 100 and 200Hz respectively. (Source: [4])

light and dark curves show the spectra of the vowel (/i/) and the unvoiced consonant (/k/) respectively. One may note the periodicity of spectrum of vowel. This is due to the harmonics of the glottal vibration superimposed over the resonant spectrum of the vocal tract in the log scale. The resonances of vocal tract give rise to broad major peaks (known as $f$ormants) in the spectrum. There is no periodicity in the spectrum of the unvoiced consonant (/k/) because the source of excitation is aperiodic in nature.

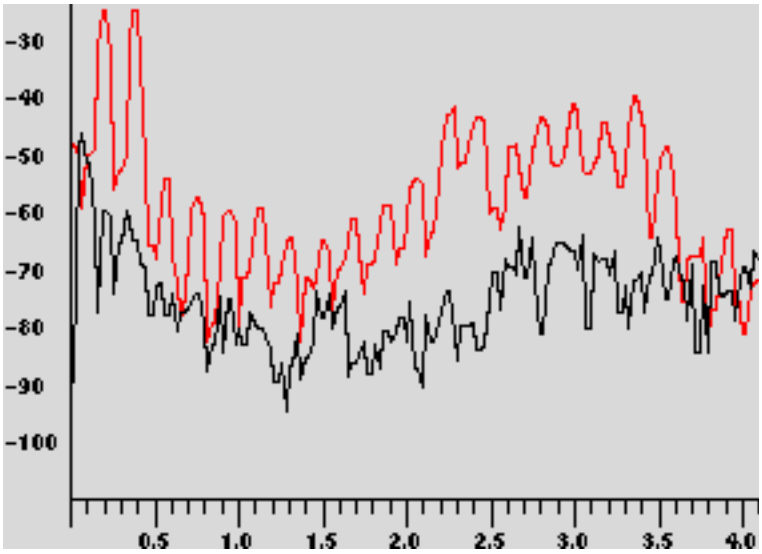

Figure 4: Power spectrum (on log scale) of a vowel (light curve) and an unvoiced sound (dark curve).

(red curve) (black curve)

2.2 Sources of variability of speech sounds

Unlike printed text, there are no well defined boundaries between phonemes or words due to flowing nature of continuous speech. For instance, an utterance of "six sheep" can easily be confused as "sick sheep" in the absence of additional information. Also, in printed text, multiple occurrences of a letter appears exactly same. In contrast, spectral and temporal characteristics of a speech sound varies a lot depending on a number of factors. Some of the important factors are explained below.

- **Physiological**: As illustrated in Fig. 3, speech waveforms of a vowel may vary due to different pitch frequencies. Also, different dimensions of vocal tract (size of head) changes the resonance frequencies of oral cavity. The resonance frequencies of male adults will, in general, be smaller than those of females, which in turn will be smaller than those of children. Thus, even if the pitch of two individuals are the same, the speech spectra can differ due to different head sizes.
- **Behavioural**: The speaking rate of people vary a lot. Syntax and semantics influence the prosodic pattern of an utterance. The accent and usage of words depend on regional and social background of a speaker. Pronunciation of unfamiliar words can deviate from the standard as shown in Fig. 6: the word "Tiruvananthapuram" is mispronounced as "Tiruvanthpuram". Such speech disfluencies aggravates an already challenging task of ASR.
- **Transducer/channel**: A microphone converts mechanical wave into an electrical signal. This transduction process may not be linear in all microphones. Fig. 5 shows variation in transduction characteristics of different telephone sets. Notice that the variation in transduction can be as much as 20dB at 3400Hz. Such a wide range of distortion modifies the spectral characteristics of speech sounds depending on the handset. In addition, compression techniques employed in mobile communication introduces additional distortion and variability.

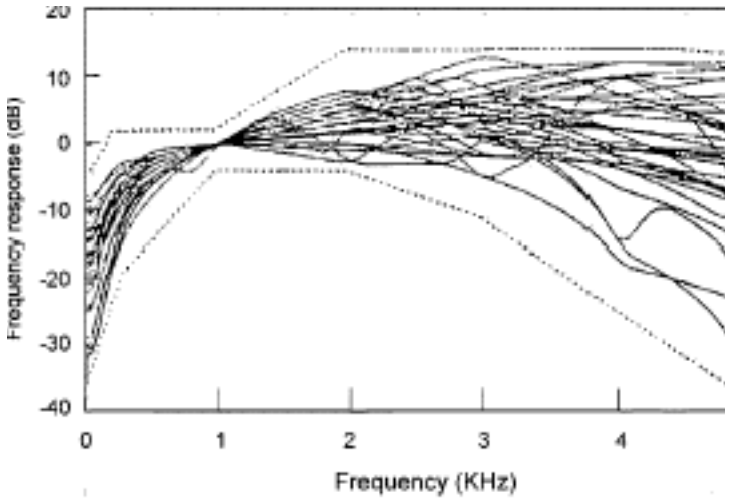

Figure 5: Diversity of transducer characteristics in telephone set The amplitude variation in transduction can be as much as 20dB (at 3400Hz) (Source: [5])

- **Environmental conditions**: Presence of background noise reduces signal to noise ratio. Background speech of neighbours gives rise to significant confusions among speech sounds. Speech recorded by a desktop speakerphone not only captures speakers voice but also multiple echos from walls and other reflecting surfaces.
- **Phonetic context**: The acoustic manifestation of a speech sound depends a lot on the preceding and following sound. This is due to inertia of articulators and is known as co-articulation. In order to illustrate this variability, Fig. 6 shows the waveform and spectrogram of a word "Tiruvananthapuram". Spectrogram shows the variation of power spectrum as a function of time. Time is along the x-axis and the frequency (0-4kHz) is along the y-axis. The power is indicated by darkness; darker a region, higher the intensity. In the figure, we can see dark bands running nearly horizontally. These correspond to vocal tract resonances (also known as $f$ormants) that change with time corresponding to different speech sounds.

  The influences of preceding and following sounds on two occurrences of the vowel /u/ within a word in Fig. 6 illustrates the variability due to phonetic context. In the first instance (at 0.15-0.22sec), the second formant *decreases* from about 1800Hz to about 1200Hz because, the vowel /u/ is preceded by /r/ whose formant frequency is higher, and followed by /w/ whose formant frequency is lower. In contrast, at 0.62-0.67sec, the second formant of /u/ *increases* from about 1200Hz to about 1500Hz because /u/ is preceded by /p/, and followed by /r/. Similar context dependent variability in temporal trajectory of formant can be observed in case of /a/ in the same figure.

Context dependent variability of speech sounds is systematic, and hence can be modelled by employing detailed phonetic units for recognition. However, other sources of variabilities have to be handled on a case by case basis. Now, let us see how such a wide variety of variabilities are handled by an ASR system.

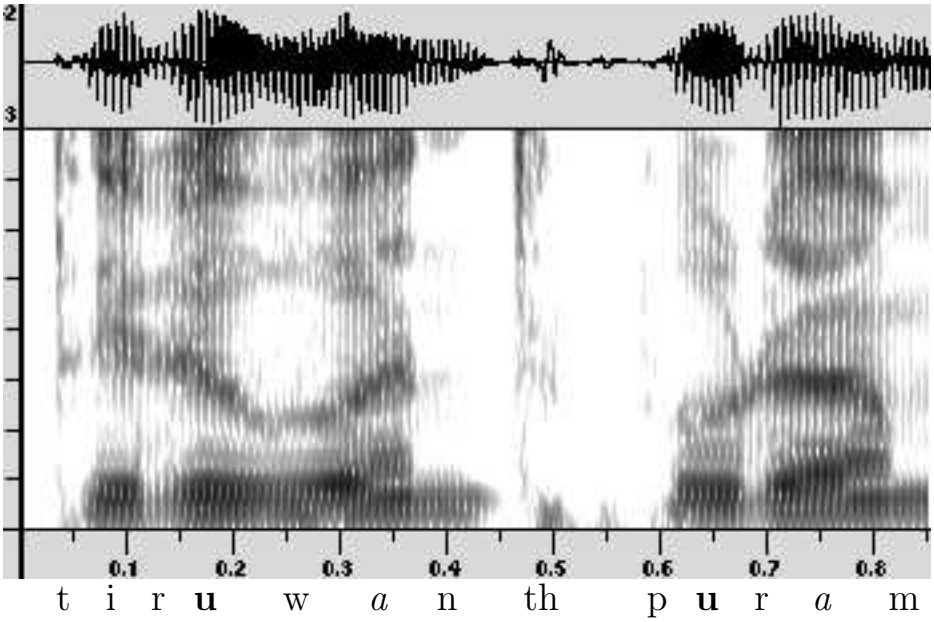

Figure 6: Spectrogram of a word "Tiruvananthapuram". Firstly, it is mispronounced as "Tiruvanthpuram". Also, notice the phonetic context dependent differences in formant trajectories of two occurrences of the same vowel /u/. In the first case (at 0.15-0.22sec), the second formant is decreasing whereas in the second case (at 0.62-0.67sec), it is increasing slowly.

3 How speech is recognized?

Speech recognition is a special case of pattern recognition. Figure 7 shows the processing stages involved in speech recognition. There are two phases in supervised pattern recognition, viz., training and testing. The process of extraction of features relevant for classification is common to both phases. During the training phase, the parameters of the classification model are estimated using a large number of class exemplars (training data). During the testing or recognition phase, the features of a test pattern (test speech data) are matched with the trained model of each and every class. The test pattern is declared to belong to that class whose model matches the test pattern best.

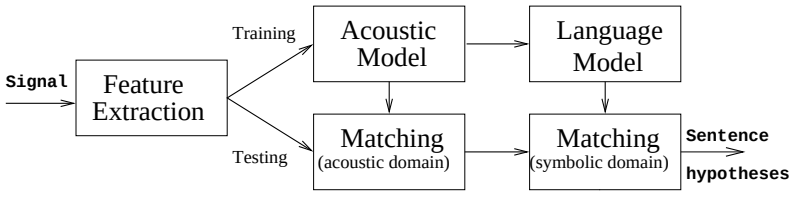

Figure 7: A block diagram of a typical speech recognition system.

The goal of speech recognition is to generate the optimal word sequence subject to linguistic constraints.

The sentence is composed of linugistic units such as words, syllables, phonemes. The acoustic evidence provided by the acoustic models of such units is combined with the rules of constructing valid and meaningful sentences in the language to hypothesise the sentence. Therefore, in case of speech recognition, the pattern matching stage can be viewed as taking place in two domains: acoustic and symbolic. In the acoustic domain, a feature vector corresponding to a small segment of test speech (called a frame of speech) is matched with the acoustic model of each and every class. The segment is assigned a set of well matching class labels along with their matching scores. This process of label assignment is repeated for every feature vector in the feature vector sequence computed from the test data. The resultant lattice of label hypotheses is processed in conjunction with the language model to yield the recognised sentence.

3.1 Signal processing

The input speech signal needs to be processed to extract features relevant for recognition. This stage is common to both training and test phases. The features should aid in discriminating similar sounds and the number of features should be small so as to rein-in computation load to a manageable level. The speech waveform is blocked into segments called frames of size about 25msec and a set of features (equivalently a multi-demensional feature vector) are extracted from each frame. The time shift between successive overlapping frames is typically 10msec.

As we discussed in section 2.1, sounds are characterized by resonances of the oral tract. Hence, the features extracted from speech signal should represent the gross shape of the spectrum while ignoring fine features of spectrum such as pitch peaks as illustrated in Fig. 3. If we view the log power spectrum as a composite signal resulting from the superposition of a slowly varying component (formant structure) and a rapidly varying component (pitch harmonics), the spectral envelope (formant structure) can be retrieved by low pass filtering the log power spectrum. The reciprocal domain of log power spectrum is called cepstrum, and the coefficients of cepstrum are called cepstral coefficients.

The amplitude compression achieved by the logarithm operation is similar to the cube root amplitude compression done by human auditory system. Also, the cochlea, in the inner ear, performs filter-bank analysis and sends nerve impulses to brain that interprets different sounds. The cochlea can resolve two frequency components, played one after another, only if the components are separated by less than a quantity called 1 bark. The width of such a 'critial' band varies with frequency; it is linear upto about 1kHz and logarithmic beyond. Such a non-linear scale is also called *m*el scale. Thus, the human ear gives more importance in resolving lower frequency components than higher ones. Since such a processing is the result of natural selection, similar filter-bank analysis should yield in better speech recognition by machines. Most ASR systems perform mel scale filter bank analysis and derive cepstral coefficients called Mel Scale Filter Coefficients (MFCC). Mel filters are approximated as overlapping triangular filters of bandwidth 1 bark and spaced at 1 bark each [6]. The steps of computing MFCCs are shown in Figure 8.

Most speech recognition systems use about 12 first cepstral coefficients (cep[q], q=1,2,..12). In addition, their time derivatives (called delta cepstral coefficients) and second time derivatives (delta-delta) are also used as features for representing dynamics of speech sounds. The latter are very useful in characterising different phonetic contexts of the same sound as illustrated by the second formant trajectories of /u/ in Fig. 6. Processing of each second of speech typically generate a sequence of 100 feature vectors.

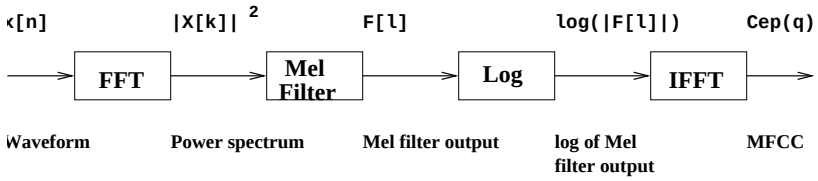

Figure 8: Stages of computation of Mel Frequency Cepstral Coefficients.

3.2 Classification of static patterns

During the training phase, the parameters of models representing phonemes or their composities need to be estimated. Due to various sources of variabilities listed in section 2.2, there is considerable overlap between phonemes in the feature space. Hence, probabilistic models of classification needs to be used. One such model is a multi-variate Gaussian Distribution: $N(\mu; \mathbf{\Sigma})$ where μ and $\mathbf{\Sigma}$ denote mean vector and Covariance matrix respectively. The likelihood of a d-dimensional vector x is computed as

$$p(\mathbf{x}|\mathbf{N}(\mu, \mathbf{\Sigma})) \;=\; \frac{\mathbf{1}}{\sqrt{(\mathbf{2}\pi)^{\mathbf{d}}|\mathbf{\Sigma}|}}\mathbf{exp}\left[-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)\right]$$

A test vector, $\mathbf{x}$ is said to belong to k^{th} class if the probability of k^{th} model generating the test pattern is the highest.

$$\mathbf{x} \in C_k \quad if p(\mathbf{x}|N(\mu_k, \mathbf{\Sigma}_k)) \;>\; p(\mathbf{x}|N(\mu_j, \mathbf{\Sigma}_j)) \quad \forall j$$

The probability distributions of features of speech sounds are not always unimodal and symmetric. Since a mixture of Gaussians can model arbitrary distributions, they are the most popularly used models of a basic unit of sound. Then, probability of matching a test vector is computed as a weighted sum of likelihoods of Gaussian densities.

$$p(\mathbf{x}) = \sum_{\mathbf{m=1}}^{\mathbf{M}} \mathbf{w_m N}(\mu_\mathbf{m}, \mathbf{\Sigma_m}) \tag{2}$$

Here w_m is the weight of the m^{th} Gaussian component and w_ms add upto 1.

3.3 Matching of temporal patterns

Speech is a temporal signal. Hence, ASR involves matching pattern sequences. When different speakers utter a given word, durations of utterances generally differ. This may be due to different speaking rates and styles of pronunciation. Thus the lengths of feature vector sequences corresponding to different repetitions of a word generally differ. Normalising the duration of utterances to a pre-specified length does not solve the problem completely due to speaking rate variations within a word. This calls for non-linear warping of speech data. Dynamic Time Warping (DTW) is a technique of finding optimal non-linear warping of test patterns so as to obtain good match with a reference pattern (model) with a reasonable computational load [7].

Although the DTW technique facilitates efficient matching of word patterns, it does not easily scale up for recognition of continuous speech. Firstly, DTW uses a deterministic reference template. Secondly, the boundary between adjacent words in a sentence is fuzzy. Thirdly, it is difficult to adequately train models of all words of a language. Also, new words are added on a daily basis. Hence, it is imperative to use a probabilistic method of modeling and matching sequences; a method that also permits generation of a model, for a new word, composed from well trained models of smaller linguistic units.

3.4 Model of continuous speech recognition

Given a trained speech recognition model and a test speech signal, the goal is to hypothesise the best sentence-a word sequence. If $\mathbf{A}$ represents the acoustic feature sequence extracted from the test data, the speech recognition system should yield the optimal word sequence, $\widehat{\mathbf{W}}$, which matches $\mathbf{A}$ best.

$$\widehat{\mathbf{W}} = \underset{\mathbf{W}}{\text{argmax}} \quad P(\mathbf{W}|\mathbf{A})$$

Re-arrangement of the above equation yields Bayes' rule:

$$P(\mathbf{W}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{A})}$$

Here, $P(\mathbf{A}|\mathbf{W})$ is the likelihood of the feature sequence $\mathbf{A}$ given the acoustic model of the word sequence, $\mathbf{W}$. $P(\mathbf{W})$ is the probability of the word sequence; this is computed from the language model. P($\mathbf{A}$) is the *a priori* probability of the feature sequence; it is independent of acoustic and language model, and can be ignored in the maximisation operation. Thus, the probability of a word sequence is approximated as the product of the probabilities of the acoustic model, $P(\mathbf{A}|\mathbf{W})$, and that of the language model, $P(\mathbf{W})$.

3.5 Acoustic model

Hidden Markov Models (HMMs) are the most popular models used in the area of continuous speech recognition. HMMs are capable of modeling and matching sequences that have inherent variability in length as well as acoustic characteristics. Here, we will introduce the basic elements of HMM and refer the readers to standard textbooks [8] and tutorials [9, 2] for details.

HMM represents a temporal pattern in the form of a Finite State Network (FSN). Each state models spectral characteristics of a quasi-stationary segment of speech. At every time instant (frame of speech), the system either continues to stay in a state or makes a transition to another in a probabilistic manner. The ij^{th} element of the transition matrix, a_{ij}, denotes the probability of transition from i^{th} state to j^{th} state. The expected duration of a state is $1/(1 - a_{ii})$; thus, a_{ii} models the intrinsic duration of a state. HMM is a doubly stochastic model. The probability distribution, $p(\mathbf{x})$, associated with a state gives the likelihood of a feature vector $\mathbf{x}$ belonging to that state. The distribution can be a GMM (Eqn. 2). Such a versatile probability distribution models highly variable nature of speech sounds.

The popularity of HMM is due to the existence of efficient algorithms for estimation of parameters of the model ($\{a_{ii}\}$, μ, $\mathbf{\Sigma}$) from the training data, and due to efficient algorithms for recognition [9]. Another advantage of HMM is its ability to integrate language models as well as explained in Section 3.6. In the following, we illustrate these concepts with an example.

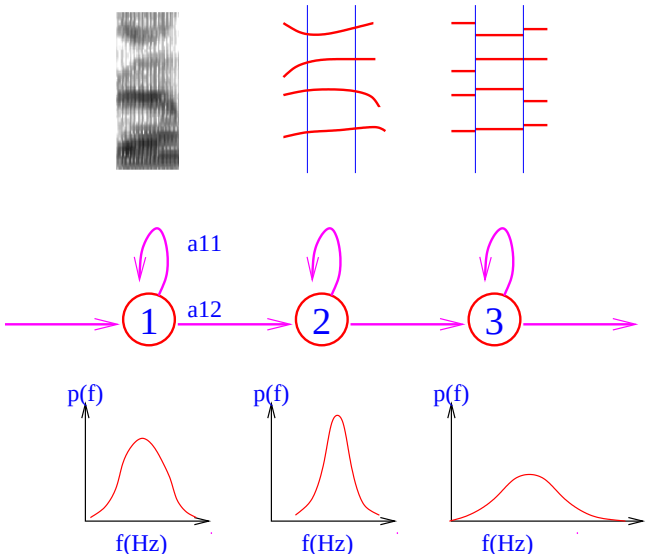

Figure 9: The trajectories of first 4 formants of the vowel /a/ occurring at 0.75sec in the Figure 6 and a HMM for vowel /a/. The left diagram of the top panel shows the spectrogram, the middle diagram is the schematic trajectory and the rightmost diagram shows a quasi-stationary approximation. The 3 states of the HMM corresponding to 3 segments of the vowel are shown in the middle panel. The bottom panel shows the Gaussian distributions associated with the states.

The 3 distributions represent variation of one feature (2nd formant) in the 3 states. The 2nd formant varies a lot in the 3rd segment; hence the 3rd Gaussian is 'wide' (has a large variance).

Let us start with the spectrogram shown in Figure 6. Let us assume that the formant frequencies are the features representing the speech sounds. Figure 9 shows the first 4 formants trajectories of the vowel /a/ occurring at the interval 0.7-0.8sec and a HMM for the vowel. In the top panel, the spectrogram is shown at left. The middle diagram is a schematic of the trajectories of 4 formants of the vowel. It can be seen that the middle portion of the formant trajectories of the vowel /a/ do not vary much compared to the left and right segments. The formants in the latter segments vary depending on the neighbouring phonemes. Thus it is desirable to represent the left, middle and right segments of a phoneme by separate probabilistic models. Such a quasi-stationary approximation of the formant trajectories is shown in the right diagram of the top panel of Figure 9. Each such segment is represented by a **state** of the vowel HMM. The 3 states corresponding to the 3 segments are shown in the middle panel of the Figure. Each state is associated with a probability distribution. The bottom panel of the Figure shows the Gaussian distributions corresponding to the 3 states. The mean values, μ_i, of the 3 states are different reflecting the average values of formant frequencies in the 3 segments. It may be noted that the variance of the Gaussian of middle state is less than those of the other states. The HMM shown in the Fig. is called a left-to-right model wherein transition from a state is permitted to a state on it's right (but not to a state on left). This reflects the causal nature of speech signal.

3.6 Language model

Given a sequence of test feature vectors, one can compute the likelihood of each phoneme model generating each frame of speech. Subsequently, one can generate, using Viterbi algorithm, a most likely phone sequence or a lattice of phone hypotheses. The role of the language model is to derive the best sentence hypothesis subject to constraints of the language. The language model incorporates various types of linguistic information. The lexicon specifies the sequences of phonemes which form valid words of the language. The syntax describes the rules of combining words to form valid sentences.

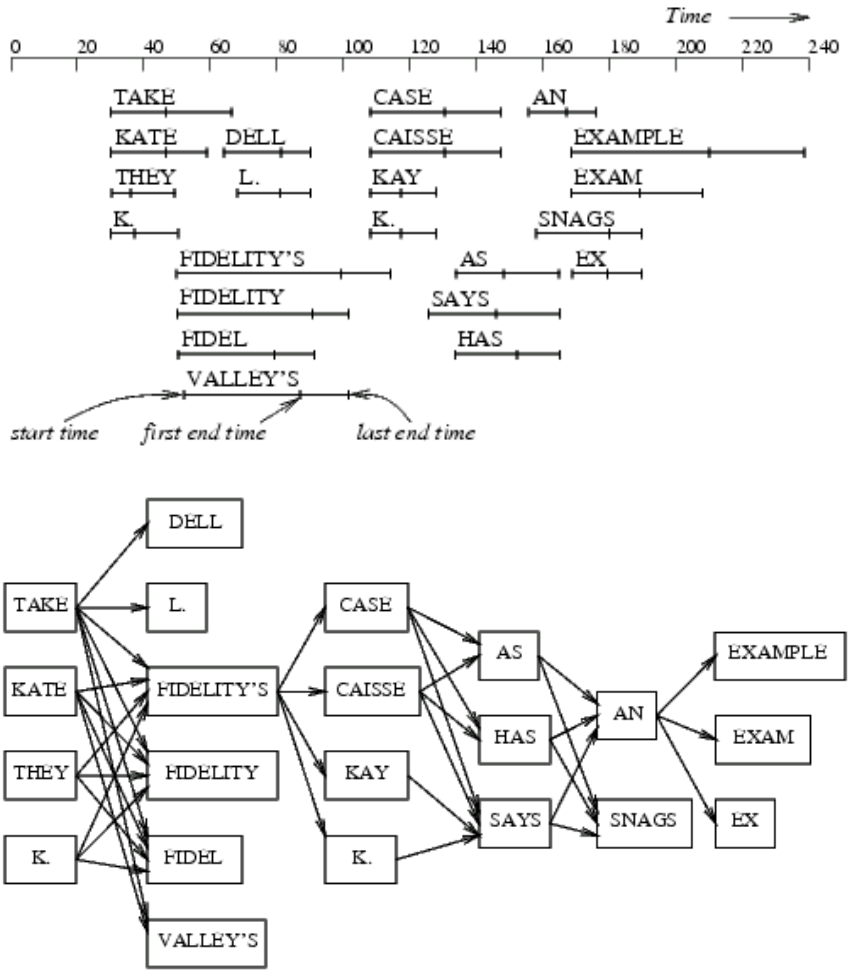

Figure 10: A lattice of word hypotheses generated for utterance: "Take Fidelity's as an example". The lower panel shows the lattice in the form of a DAG (source [10]).

An example of word hypotheses generated by an ASR based on acoustic evidence and lexical knowledge is shown in Fig. 10. Each word hypothesis is assigned time-stamps: begin, earliest and latest end points of the word. For instance, the word "Take" can correspond to any segement starting at 28th frame and ending anywhere between 46th and 68th frame. As shown in the lower panel of the figure, a Directed

Acyclic Graph can be derived based on frame continuity constraints and word-pair grammar. Sentence hypothesis can be refined by using sophisticated language models.

Two main categories of language models are statistical models and word transition network. The simplest of statistical grammars is the N-gram grammar. It specifies the probability of the n^{th} word in a sequence, given the previous $n-1$ words of the sequence.

$$p(w_i \mid w_{i-1}, w_{i-2}, \cdots, w_1) \;=\; p(w_i \mid w_{i-1}, w_{i-2}, \cdots, w_{i-N+1})$$

Bigram (N=2) and trigram (N=3) are prevalent forms of N-gram grammar. The N-gram grammar can be applied to other characteristics of words such as parts of speech.

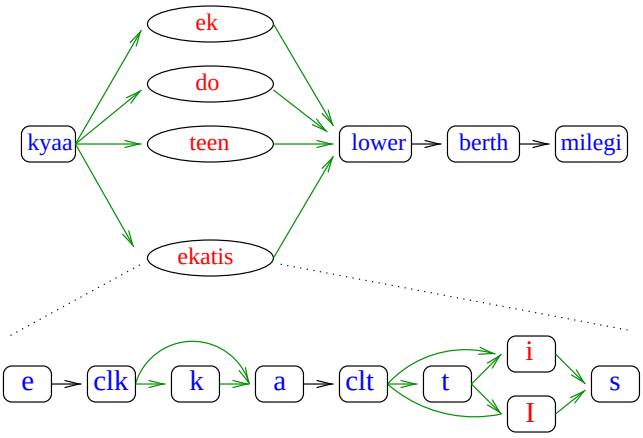

Figure 11: A word transition network employed in a Hindi speech recognition system. HMM of each word is composed of HMMs of the constituent phone-like units as shown for the word एकतीस.

Word transition network is a popular language model for speech recognition in a well-defined task domain. It incorporates both syntactic and semantic constraints of the task domain to achieve better performance. Figure 11 shows a fragment of word transition net of a Hindi speech recognition system in the context of railway reservation enquiry task [11]. The top panel of the Figure shows a word net which represents a set of 31 valid sentences of the task domain. The word net can be represented as a super HMM comprising of word HMMs; each word HMM is, in turn, composed of HMMs of the phone-like units. The architecture of a composite word HMM, shown in the bottom panel, allows for phone substitution and deletion—a phenomena common in fluent speech. The topology of the composite word HMM shown in the figure permits 8 different pronunciations of the word एकतीस.

The best sentence is hypothesised after a judicious combination of evidences from acoustic and language models:

$$P(\mathbf{W}|\mathbf{A}) = P(\mathbf{A}|\mathbf{W}) \;\; P(\mathbf{W})^{\gamma}$$

Here, γ is the weight factor of the language model. If necessary, multiple sentence hypotheses with associated scores can be derived from the word lattice (DAG). For speech understanding task, quite often, identification of keywords is sufficient.

4 Speech recognition systems

Application systems using speech recognition technology range from isolated word recognition systems used in toys to speaker independent, fluent conversation over telephone with information dissemination systems. Just a small sample is given here.

4.1 International scene

There are several Dictation Systems of very large (over 100,000) vocabulary with recognition accuracy of about 97%; these are speaker dependent and needs to be trained. Automatic Call Routing systems based on ASR technology have been in use for more than a decade. A few airlines employ speech recognition systems to provide information to callers and even permit booking of airline ticket. An ASR system recognises and understands quries in fluently spoken German, and provides train timetable information [12]. Other example ASR applications are automated switchboard of a large German company, the movie or football information system [13].

4.2 Indian scene

Commercial ASR based systems have started appearing in India. Several dictation systems trained with English of Indian accent are in the market. Mobile phones permit name dialling, i.e., calling a person by speaking his name. Other speaker independent, narrowband ASR systems follow a machine-guided, menu driven mode. The caller is asked to provide one piece of information at a time; the ASR systems used are of isolated 'word' recognition type. Indian Railways provides information about departure/arrival information of trains on speaking the name of the train over telephone. Similar Isolated 'Word' Recognition system deployed by IRCTC [14] enables a mobile user to make train reservations. Proposals have been made to develop and implement ASR systems that provide information from the online database of Indian Railway to a caller who can query the system in terms of fluently spoken sentences in Indian Languages [15].

5 What needs to be done?

Although HMM is the prevalent model, it has several inherent drawbacks. HMM models speech signal as a sequence of probability distributions based on the premise that speech is a quasi-stationarity signal; this premise is not valid in case of dynamic sounds such as plosives or glides. The standard training and decoding algorithms make first order Markovian assumption: the feature vector depends only on the current state. This is in variance with context dependent properties of phonemes. Probability distributions associated with states assume that successive feature vectors are uncorrelated, whereas these *a*re correlated because speech is generated by a natural system with associated inertia of articulators. The state duration distribution is exponential with mode at zero; this is in contrast to non-zero durations of speech sounds.

The need for sophisticated language models is even more accute. Humans can cope with speech far from ideal: low SNR due to strong background noise (ex: car moving in a highway); transient noise such as

cough, telephone rings; reverberant speech from speakerphone or in a hall; speech disfluencies such as "ah"s, "am" or repetitions, ungrammatical sequence of words etc. The human auditory system, especially the brain, has evolved over millennia to 'model' and handle a wide variety of distortions. Adavanced models of natural language needs to developed that mimic the language capabilities of human brain. They should be capable of integrating not only syntactic and semantic knowledge, but also pragmatics. Thus, better models are needed for developing ASR systems similar to those portrayed in science fiction movies.

ASR systems are not perfect. Versatile systems automatically detect when they are not doing a good job of recognition, and transfer control to human operator. There have been attempts to embed hand held devices such as mobile phones using distributed ASR to conserve bandwidth. Here, the device extracts features from the speech waveform and transmits them to a server which carries out the speech recognition task.

5.1 Indian context

Popular ASR systems are based on statistical models. Robust estimation of parameters of the model needs a large amount of annotated phonetically and prosodically rich speech data. speech has to recorded large number of persons diverse age group speaking in a variety of real-life situations. Some progress is made in developing labelled spoken corpus for Indian languages [16] and a lot needs to be done.

Several organizations in India are engaged in Research and Development of ASR systems [17]. Courses in Speech Technology and conferences in the area of speech and language are held regularly [18]. Quite a projects are being sponsored by the Technology Development of Indian Languages Programme of Government of India [19].

6 Conclusions

Significant progress has been made in the past couple of decades in the area of spoken language technology. This has led to deployment of speech recognition systems in a few application domains. Yet, the current engineering models of speech and language do not adequately model the natural language capabilities of human brain. The cognitive aspects of human brain are complex and development of appropriate models is still a challenging research task. Such a development will lead to Ubiquitous Speech Communication Interfaces through which people will be able to interact with machines as conveniently and naturally as do amongst themselves.

7 References

1. B.H. Juang, and S. Furui, "Automatic Recognition and Understanding of Spoken Language–A First Step Toward Natural HumanMachine Communication, Proc. IEEE, **88**, No. 8, 2000, pp. 1142-1165.

2. http://htk.eng.cam.ac.uk/docs/docs.shtml

3. http://murray.newcastle.edu.au/users/staff/speech/ home_pages/tutorial_acoustic.html

4. http://www.haskins.yale.edu/haskins/ HEADS/MMSP/acoustic.html

5. H. C. Wang, M.-S. Chen, and T. Yang, "A novel approach to the speaker identification over telephone networks," in Proc. ICASSP-93. 1993, vol. 2, pp. 407410.

6. Davis S and Mermelstein P, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on ASSP, vol. **28**, pp. 357-366.

7. Hiroaki Sakoe and Seibi Chiba, "Dynamic Programming Algorithms Optimization for Spoken Word Recoginition", IEEE Trans on acoustics, speech and signal processing, vol.ASSP-26, no.1, december 1978.

8. L.R.Rabiner and B.H.Juang, "Fundamentals of Speech Recognition", Prentice Hall, New Jersey, 1993.

9. Rabiner L R, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" Proc. IEEE, vol. **77**, 1989, pp. 257-286.

10. M.K.Ravishankar, "Efficient algorithms for Speech Recognition", PhD thesis: CMU-CS-96-143.

11. Samudravijaya K, "Hindi Speech Recognition", J. Acoust. Soc. India, **29**(1), pp. 385-393, 2001.

12. H.Aust, M.Oerder, F.Seide and V.Steinbiss "The Philips automatic train timetable information system", Speech Commun. **17** (1995) 249-262.

13. E.Noth and A.Homdasch, "Experiences with Commercial Telephone-based Dialogue Systems", IT Information Technology, **46** (2004) 6, pp.315-321.

14. http://www.irctc.co.in/voice.html

15. http://www.au-kbc.org/dfki/

16. Samudravijaya K, Rao P V S and Agrawal S S, "Hindi speech database", in Proc. Int. Conf. Spoken Language Processing ICSLP00, Beijing, 2000; CDROM: 00192.pdf.

17. http://speech.tifr.res.in/slp_in_india.html

18. http://ltrc.iiit.net/icon2005; http://www.ncst.ernet.in/kbcs2004

19. http://tdil.mit.gov.in