

Speech and Speaker Recognition: A Tutorial

Samudravijaya K

Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005

Email: chief@tifr.res.in

Abstract

Speech is the primary mode of communication among human beings. So, it has the potential of being an important mode of interaction with computers. In this tutorial, salient features of speech and speaker recognition systems are introduced. Speech recognition systems permit ordinary people to speak to the computer to retrieve information. Speaker recognition systems provides a convenient means of establishing the identify of a person based on his voice. Issues related to speech signal processing and training acoustic as well as language models are covered. The basic blocks of a Hindi speech recognition system and a text prompted speaker verification system are also described.

1 Introduction

A convenient and user-friendly interface for Human Computer Interaction is an important technological issue. The prevalent computer interface is via a keyboard or a pointing device for input and a visual display unit or a printer for output. In the current Indian context, these machine-oriented interfaces restrict the computer usage to a minuscule fraction of the population, who are both computer literate and conversant with written English. Communication among human beings is dominated by spoken language. Therefore, it is natural for people to expect speech interfaces with computers. Computers which can speak and recognise speech in native language enable even a common man to reap the benefit of information technology. This tutorial deals with two aspects of speech input, namely speech recognition and speaker recognition.

Machine recognition of speech involves generating a sequence of words which best matches the given speech signal. In the speaker independent mode of speech recognition, the computer should ignore the speaker specific characteristics of the speech signal and extract the intended message. On the other hand, in case of speaker recognition, the machine should extract speaker characteristics in the acoustic signal. Hence, the focus of signal processing are, in a sense, complimentary in cases of speech and speaker recognition.

This tutorial is organised as follows. In the next section, production of speech sounds

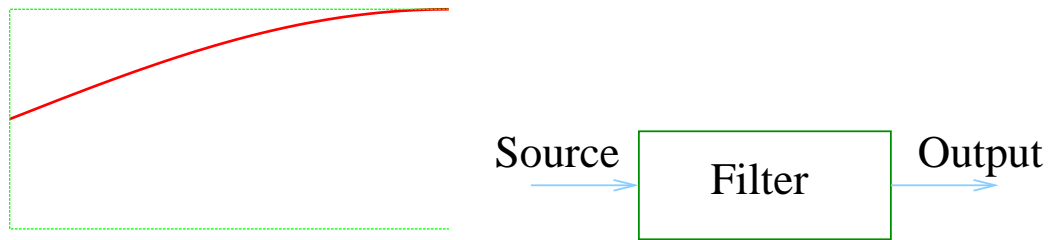


Figure 1: The diagram on left is a schematic of the uniform tube model of oral tract. The dark line indicates the amplitude of vibration of air molecules during resonance as a function of position along the vocal tract: glottis (left end) through lips (right end). The right diagram is an engineering model of the speech production. Here, vocal cavity is the filter.

and signal processing aspects common to speech and speaker recognition are covered. In section 3, the basic concepts of pattern recognition as well as the challenges and approaches to speech recognition are described. Section 4 deals with speaker recognition. Some concluding remarks are drawn in section 5.

2 Speech production and signal processing

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.

Speech sounds are generated when the vocal tract, the air passage from the glottis to the lips, is excited. The mode of excitation can be of 3 types: periodic, aperiodic and mixed. In case of voiced sounds such as vowels, the excitation is periodic. The periodic opening and closing of glottis results in puffs of air exciting vocal tract. In case of generation of neutral vowel /a/, the vocal tract can be approximated, during the closed phase of glottis vibration, as a uniform tube closed at one end as shown in the left diagram of Figure 1. 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

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

The frequencies of harmonics will be 1500Hz, 2500Hz etc. Thus, we should expect peaks in the frequency spectrum of the vowel /a/ at these frequencies. These peaks in the

spectrum, due to resonances in the vocal tract, are called **formants**. Different speech sounds are generated by changing the size and shape of the resonant cavity resulting in different values of frequency, amplitude and bandwidth of formants.

The source of periodic excitation, the glottis, vibrates with a frequency popularly known as the pitch frequency, which ranges from 75 to 300Hz. The frequency spectrum of glottal vibration is a sequence of impulses of diminishing amplitudes, the first corresponding to the fundamental frequency and the rest to its harmonics.

As the source of excitation and vocal tract shapes are relatively independent, one can model them separately. The right diagram of Figure 1 shows the **source-filter model** of speech production. In this model, a time-varying filter represents the vocal tract which is excited by an appropriate source. The output of the filter is the speech wave, $s(n)$, which can be written as the convolution of the source, $e(n)$ and the impulse response function, $h(n)$, of the filter (vocal tract).

$$s(n) = e(n) * h(n)$$

$$\log(|S(k)|^2) = \log(|E(k)|^2) + \log(|H(k)|^2) \quad (2)$$

In the frequency domain, the log power spectrum of speech is the sum of the log power spectra of source and filter.

Figure 2 shows typical power spectra of two speech sounds of the Hindi word “ki” on a log scale. The 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 as expected from the Equation 2. The resonance of vocal tract give rise to broad major peaks (formants) in the spectrum. There is no periodicity in the spectrum of the unvoiced consonant (/k/) because the source of excitation is aperiodic in nature.

Speech recognition is the process of deriving the sequence of speech sounds best matching the input speech signal. Speech sounds are characterised by the size and shape of filter (vocal cavity) which is represented by the spectrum of the filter, $H(k)$. Therefore, the source characteristics such as the fundamental frequency, signal amplitude etc. should be ignored in speech recognition. As can be seen from Equation 2, the log power spectrum of speech is the superposition of the log power spectra of source and filter. From Figure 2, one can see that the log power spectrum of source varies rapidly with frequency whereas that of filter varies slowly. Therefore, if we pass this composite log power spectrum, $\log(|S(k)|^2)$, through a low pass filter, only the characteristics of the filter (speech sound) remains. This process is called liftering. This is achieved by taking inverse Fourier transform of log power spectrum and retaining only the first few coefficients. The resultant spectrum

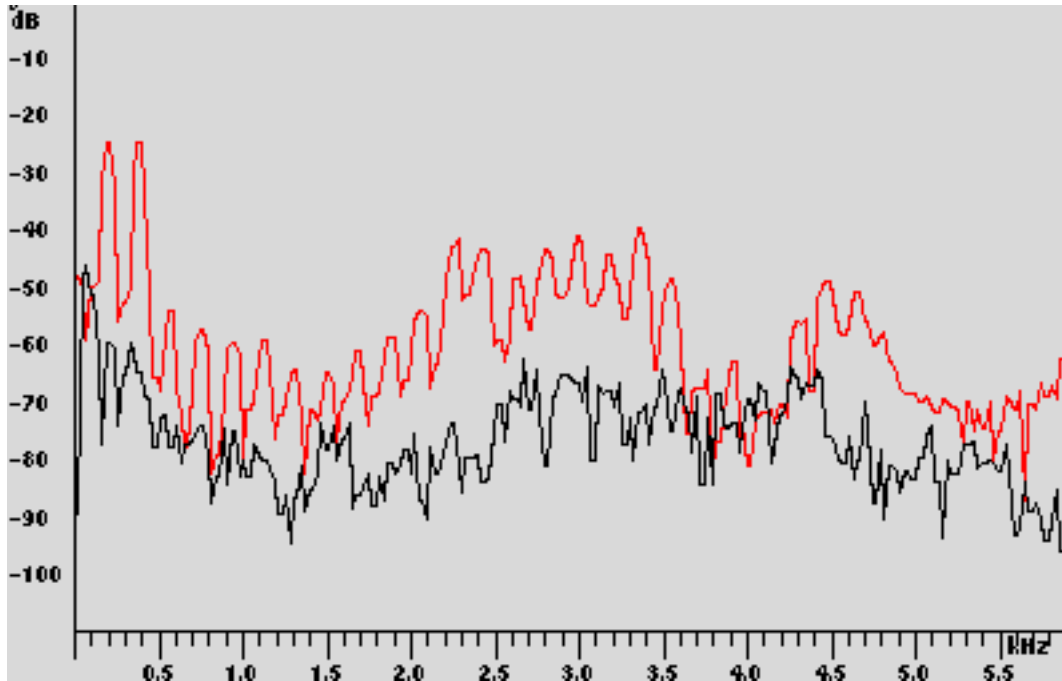


Figure 2: Power spectrum (on log scale) of a vowel and an unvoiced sound.

is called cepstrum and the coefficients, $cep(q)$, are called the cepstral coefficients.

$$cep(q) = IFFT\{\log(|S(k)|^2)\} \quad q = 0, 1, \dots, N - 1$$

Most speech recognition systems use the cepstral coefficients and their time derivatives as features for representing speech sounds. The performance of systems can be improved by emulating the signal processing in the peripheral auditory system of humans. Such enhancements include exploiting non-linear frequency sensitivity of cochlea, power-law of intensity-loudness curve etc [1].

3 Speech recognition

Speech recognition is a special case of pattern recognition. Figure 3 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

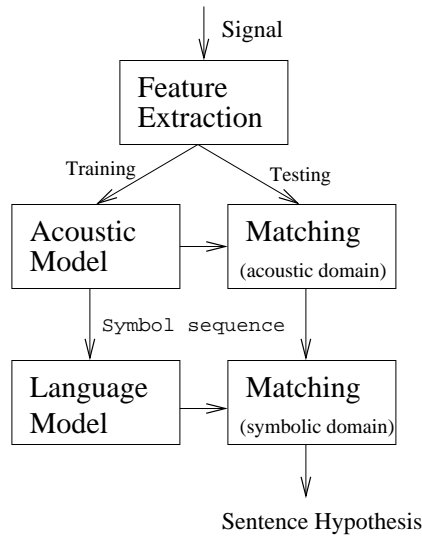


Figure 3: The block diagram of a typical speech recognition system.

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.

The goal of speech recognition is to generate the optimal word sequence subject to linguistic constraints. The sentence is composed of linguistic units such as words, syllables, phonemes. In speech recognition, a sentence model is assumed to be a sequence of models of such smaller units. 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 the label of the class with the highest matching score. This process of label assignment is repeated for every feature vector in the feature vector sequence computed from the test data. The resultant sequence of labels or a lattice of label hypotheses is processed in conjunction with the language model to yield the recognised sentence.

3.1 Classification of static patterns

Figure 4 shows the vowels of English in the F1-F2 space where F1 and F2 are the frequencies of the first two formants as measured by Peterson and Barney in a classic work [2]. It

is obvious that the vowel tokens from multiple speakers form clusters in the F1-F2 space. Since the shapes of clusters are ellipsoid, one has to use weighted Euclidean distance to measure the distance of a test vector, \mathbf{x} , to the centroid of k^{th} class, $\mu_{\mathbf{k}}$.

$$d = (\mathbf{x} - \mu_{\mathbf{k}})^2 / \Sigma$$

Here, Σ is the covariance matrix and accounts for non-spherical shapes of the clusters.

Since some of the clusters are overlapping, it is better to use probabilistic models. The simplest model is the Gaussian Distribution: $N(\mu; \sigma)$.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right\}$$

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

$$x \in C_k \text{ if } p(x|N(\mu_k; \sigma_k)) \geq p(x|N(\mu_j; \sigma_j)) \quad \forall j$$

3.2 Classification of temporal patterns

Speech is a time-varying signal. Normally short-time processing of speech is employed to extract features from speech data. Here, the speech signal is segmented into (possibly overlapping) regions, each called a frame of speech. Corresponding to a frame of speech, a feature vector such as the cepstral coefficients is computed. This process is repeated for all frames of a speech data yielding a sequence of feature vectors. Thus, speech recognition involves matching a feature vector sequence with models of words.

When different speakers utter a given word, the duration of speech samples 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 speech samples 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. The dynamic time warping 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 [3].

3.3 Why speech recognition is difficult

There are many issues which limit the performance of a speech recognition system. The main barrier is the variability of speech signal. A given word spoken by different persons

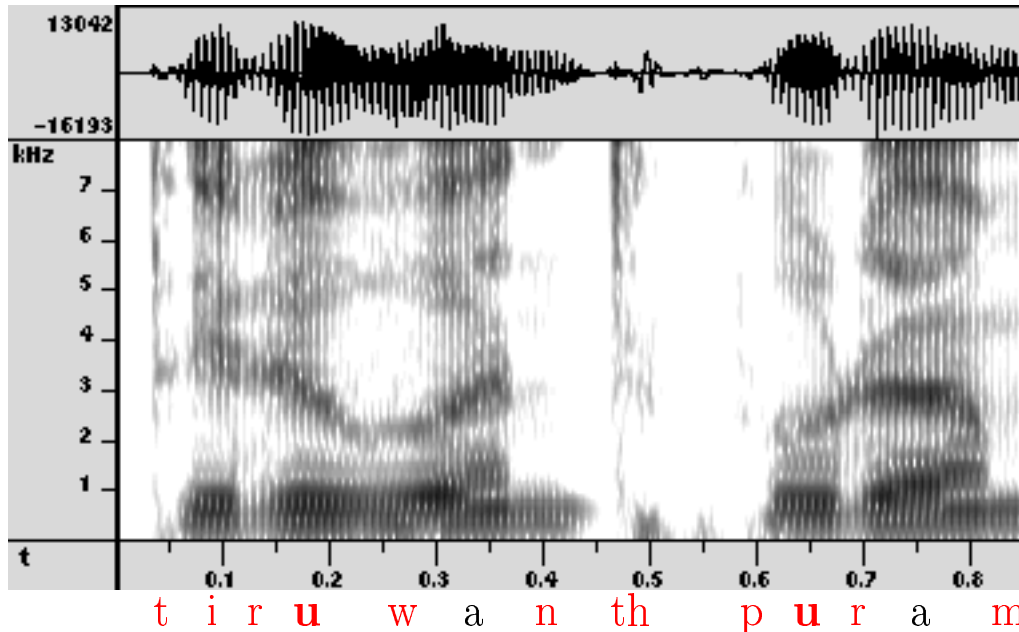


Figure 5: The time waveform and spectrogram of the word “tiruvanthapuram”. Note the phonetic context dependent variation of spectra of two tokens of /a/.

can have different spectral and temporal properties due to variations in physiological characteristics, emotional status and cultural background. For example, females, in general, have a shorter vocal tract than a male. So, the formant frequencies of a vowel spoken by female speakers are, in general, higher than those of males according to Equation 1.

For all practical purposes, speech can be adequately described in terms of linguistic units called phonemes. For example, each character of Devanagari script represents one phoneme. However, there are no well defined boundaries between phonemes in continuous speech. The spectral characteristics change continuously due to the inertia of the articulators which move from the position of one phoneme to the position of the next phoneme. Also, the articulators move to the position of the next phoneme even while the current sound is being uttered. Consequently, the acoustic properties of a speech sound not only depends on the identity of the corresponding phoneme, but also on the neighbouring sounds. This variability due to phonetic context, however, can be predicted unlike speaker dependent variations.

The effect of phonetic context on the spectra of phonemes is illustrated in Figure 5. This Figure shows the time waveform and spectrogram of the word “tiruvanthapuram”. Spectrogram is a method of displaying spectrum of a signal as a function of time. It is a projection of a function of 3 variables on a 2-dimensional plane. Here time and frequency are plotted along x and y axes respectively. The amplitude at a given frequency and at a

given time is indicated by the darkness of the corresponding pixel. The dark bands represent the temporal variation of formant frequencies. The different segments corresponding to the phonemes of the word are marked below the spectrogram for reference. A popular name for spectrogram is ‘voice print’. Just like a finger print is a characteristic of an individual, a spectrogram is a visual representation of an utterance.

First of all, note that the speaker has mispronounced the long word “tiruvanthapuram” as “tiruvanthpuram”. Human beings have no problem in translating this to the correct word. However, such articulatory laxities pose a problem for a machine. Secondly, two occurrences of the phoneme /a/ in the word have different spectral trajectories. The temporal variation of spectra of two instances of the vowel /a/ are different due to different phonemic contexts. Specifically, the second formant (F2) of the vowel is increasing with time in the first case, whereas it is nearly steady (and declines slightly later) in the case of /a/ following /r/. Similarly, the directions of movement of F3 in the two examples are opposite. Similar phonetic context dependent variation can be observed in case of two tokens of the vowel /u/ as well.

Superposition of background noise and extraneous signals in the transmission channel decrease the signal to noise ratio of speech signal. Acoustic characteristics of room, microphone and transmission channel get convolved with the speech signal. All such effects add to the variations of speech signal.

3.4 Model of 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.

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

Using Bayes’ rule, we can re-write $P(\mathbf{W}|\mathbf{A})$ as

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

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

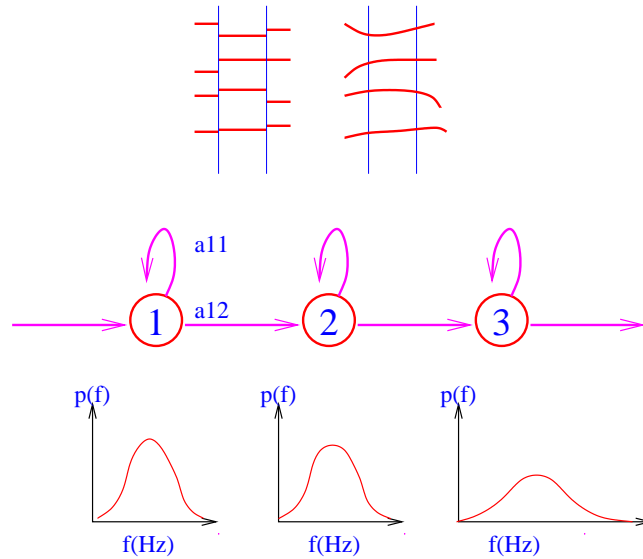


Figure 6: The top panel shows the trajectories of first 4 formants of the vowel /a/ occurring at 750msec in the Figure 5. The right diagram of the top panel is the actual trajectory and the left diagram of the top panel is a quasi-stationary approximation. The 3 states of the HMM corresponding to 3 segments of the vowel are shown in the middle panel. The bottom panel shows the Gaussian distributions associated with the states.

is independent of acoustic and language model, and can be ignored in the maximisation operation. Thus, the probability of a word sequence is 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

The acoustic model for speech recognition should be capable of modeling predictable variations (context dependent variations) of acoustic characteristics of speech sounds as well as other variations due to speaker, channel etc. A probabilistic model would be the best to account for the overlap of phonemes in the acoustic space. In addition, the model should also handle variable durations of phonemes due to prosody. Most popular acoustic model in use is the hidden Markov model (HMM). Here, we will introduce the basic elements of HMM and refer the readers to standard textbooks [4] for details.

Let us start with the spectrogram shown in Figure 5. Let us assume that the first 4 formant frequencies are the features representing the speech sounds. Consider the formant trajectories of the vowel /a/ occurring at 750msec. A schematic of the trajectories of 4

formants of this vowel is shown at the top right panel of Figure 6. It can be seen that the middle portion of the formant trajectories of the vowel /a/ does 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 left diagram of the top panel of Figure 6. Each such segment is represented by a **state** of 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.

In HMM, each phoneme is represented by a sequence of states as shown in Figure 6. The system can make transition from one state to another. Transition from a state to itself accounts for variable duration of phonemes. The probability of transition from i^{th} state to j^{th} is denoted by a_{ij} where j runs from i to N . Here, N is the number of states of the model; in this case it is 3. Gaussian distributions are characterised by mean vectors μ_i and covariance matrix \mathbf{W} . Due to laxity of articulation, it is possible that the first segment may not be pronounced. This can be modeled by skipping the first state, i.e., the system enters the second state without going through the first state. The parameter π_i represents the probability that the system jumps to the i^{th} state, skipping all previous states. The number of states of a model is predetermined by the system developer. The set of parameters $\{a_{ij}\}$, $\{\mu_i\}$, \mathbf{W} and $\{\pi_i\}$ of a model have to be estimated from the training data. The popularity of HMM is due to existence of efficient algorithms for estimation of these parameters and due to efficient recognition algorithms. Another advantage of HMM is its ability to integrate language models as well. This is explained in Section 3.6.

3.6 Language model

Given a sequence of feature vectors corresponding to given speech data, one can compute the likelihood of each phoneme model generating each frame of speech. In turn, one can generate 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. Human beings use semantics and pragmatics to recognise an utterance. Two main categories of language models are statistical models and word transition network.

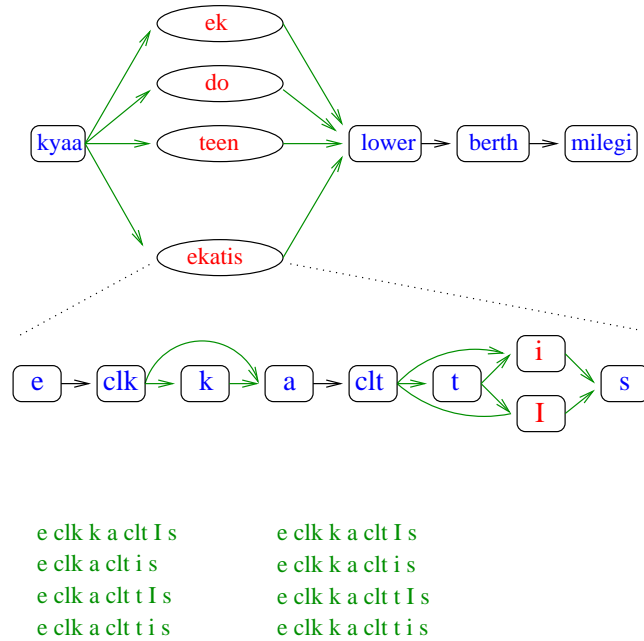


Figure 7: A word transition network employed in a Hindi speech recognition system. HMM of each word is composed of HMMs of the constituent phonemes.

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.

$$p(w_i | w_{i-1}, w_{i-2}, \dots, w_1) = p(w_i | w_{i-1}, w_{i-2}, \dots, w_{i-N+1})$$

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

Word transition network is a popular 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 7 shows a fragment of word transition net of a Hindi speech recognition system in the context of railway reservation enquiry task [5]. The top panel of the Figure shows a word net which represents 31 valid sentences of the task domain. Each word is represented by a HMM which is composed of HMMs of the phonemes of the word. The architecture of this composite HMM, shown in the middle panel, allows for phone substitution and deletion—phenomena common in fluent speech. This composite HMM permits 8 different pronunciations of the word *ekatis* which are listed in the bottom panel of the Figure.

4 Speaker recognition

Speech signal not only contains the message but also auxiliary information such as the gender/identity of the speaker, the characteristics of the room, hand set etc. In speech recognition, the aim is to extract the message while ignoring such auxiliary information. In case of speaker recognition, the goal is glean information about the speaker without much importance being given to the message.

Speaker recognition can be classified into 3 categories. Speaker identification involves determining the identity of a speaker belonging to a closed set. It is a N-way classification process. Speaker verification is a decision process of accepting or rejecting the identity claim of a person based on speech. Speaker tracking is the process of determining whether two adjacent segments of speech belongs to same speaker or not. Speaker tracking is useful in a conference call where the discussion is held between a closed set of persons.

The process of speech recognition hinges on characterising phonemes whereas the speaker recognition depends on characterising the speaker. The phonemes are characterised by resonances of vocal cavity (formants). It is not clear what are the attributes which characterise a person's voice. Humans rely more on suprasegmental features such as pitch, speaking rate. However, these can be easily mimicked and hence are not useful for applications such as speaker verification. Long term statistical features such as average spectrum do represent physiological characteristics of the vocal tract of a person. However, reliable estimation of these features need long utterances which may not be practical in many cases.

The voice quality of people are often qualified with terms such as pleasing, hoarse, resonant, breathy, nasal and so on. These are qualitative terms and are difficult to quantify. Hence, most speaker recognition systems use the same set of features which are used in speech recognition systems. In the following, we will describe one type of speaker recognition, namely speaker verification.

4.1 Speaker verification

Speaker verification is the process of validating a user's claim to an identity based on his/her voice. It is a special case of supervised pattern recognition. Figure 8 shows an outline of a typical speaker verification system. During the training phase, an acoustic model of the voice characteristics of each user is generated based on training data. During the testing phase, the characteristics of the test data is matched with the model of the claimed identity. If the matching score is above a threshold, the claim is accepted.

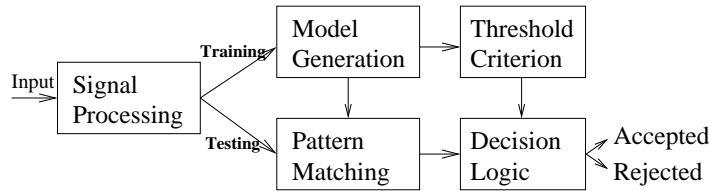


Figure 8: A block diagram of a typical speaker verification system.

Applications of speaker verification include banking transactions over telephone, access to restricted area, retrieval of information on payment basis etc. The system can make two types of errors. It may reject a claim of a genuine user. This is called False Rejection. On the other hand, it may erroneously accept the claim of an imposter; this error is called False Acceptance. The ratio of these two errors depends on the matching score threshold set acceptance. This threshold is set by the operator depending on the needs of specific application. In case of access to secure place, False Acceptance is more costly than False Rejection. On the other hand, in case of retrieval of stock tips, rejection of many genuine claims may lead to high rate of user dissatisfaction. In that case the operator may set the threshold to a low value even if this allows a few imposters to gain the information.

There are 3 modes in which speaker verification can be done. In text independent mode, the system relies only on the voice characteristics of the speaker. This mode is used in surveillance or forensic applications where there is no control over the speakers. In the text dependent mode of verification, the user is expected to say a pre-determined text—a voice password. Text dependent speaker verification system will perform better than the text independent system as the former can utilise the phonetic information of the password as well. However, this system is not yet used in large scale due to fear of ‘playback attack’. If an imposter can record the spoken password, he can gain access to system easily by playing the recorded password. To alleviate such fears, text prompted speaker verification systems have been developed. This is an improved version of the text dependent system. The difference is that the password is not pre-determined, rather a random password is generated by the system online. The user is asked to repeat the random password. If the number of distinct random passwords is large, the ‘playback attack’ is not feasible.

Figure 9 shows the outline of a text prompted speaker verification system. During testing phase, the system verifies whether the user has spoken the random text that is prompted by the system. Here, a speaker independent speech recognition system is used for better text recognition accuracy. If the speech recognition system validates the test speech data, then the system proceeds to verify the voice quality of the user. The speaker independent system is adapted to the target speaker’s model using his training data. This results in a

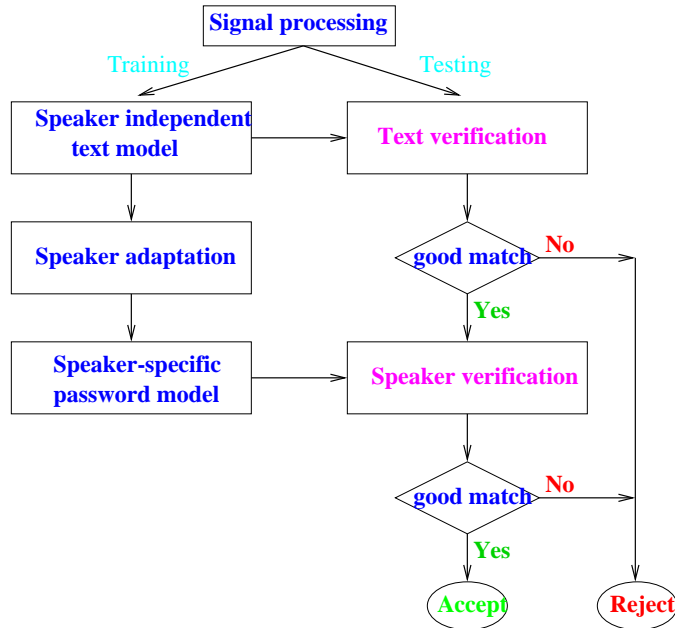


Figure 9: A block diagram of a text prompted speaker verification system.

speaker adapted phone model. The test data is again matched against this speaker-specific password model. The claim is accepted if the matching score is above a threshold.

The prompt text can be arbitrary. However, if the vocabulary is small, better adaptation of speaker independent model to a new speaker is possible. One such system prompts a 7-digit random number. There are 10 million such numbers; yet, the vocabulary of the system is just 10. Advanced versions of the systems prompt arbitrary text and use phone recognition system for validation.

5 Conclusions

Integration of computers and telecommunication system has brought the issue of convenient computer interfaces for remote access to the fore. A computer with a speech interface enables ordinary people to reap the benefit of information revolution. The ability to interact with computer in one's native language is very relevant to a multi-lingual country such as India. Hence, speech input/output systems are expected to play a major role in computer revolution in India. The development of speech interfaces involves efforts of pool of scientists and engineers well versed with speech signal and language processing in addition to language resources such as speech databases [6]. This tutorial would,

hopefully, help in the creation of such a pool of experts.

6 References

1. 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.
2. G E Peterson and H L Barney, "Control Methods Used in a Study of the Vowels", J. Acoust. Soc. Am., 24(2), pp 175-194, March 1952.
3. Hiroaki Sakoe and Seibi Chiba, "Dynamic Programming Algorithms Optimization for Spoken Word Recognition", IEEE Trans on acoustics, speech and signal processing, vol.ASSP-26, no.1, december 1978.
4. Rabiner L R, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" Proc. IEEE, vol. **77**, 1989, pp. 257-286.
5. Samudravijaya K, "Hindi Speech Recognition", J. Acoust. Soc. India, **29**(1), pp. 385-393, 2001.
6. 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.