An Introduction to Speech Technology

Dr. P. V. S. Rao Adviser, Tata Infotech Ltd.

1. Introduction

There have been two major problems with use of computers by the 'uninitiated'; the need for keyboard skills and the restraint that communication has to be tightly structured, not like informal conversation as between two individuals. Using speech for man-machine communication obviates these problems and has been a long time dream for many.

Research in speech recognition by computers has been on for over 40 years. Speech recognition has been an active area for research in India also for equally long. There is however, a long way still to go. This is because the problem is complex. For instance, an utterance might have significantly different properties even if it is spoken by the same speaker at different times – depending on context, mood and physical state of the speaker. Variability increases across speakers, depending on age, sex, region and cultural background. It is difficult for a speech recognition system to accommodate such wide variance.

Also, speech is not produced in the manner words are put out in a typewriter, letter by letter, as discrete units. In fact, the sounds of speech are generated in a continuous process, each running into the other, the way individual letters do in cursive script. The process is more like the graceful movements of a ballet dancer rather than the staccato movements of a cartoon character. Consequently, the properties of each element are significantly influenced by context – the identity of (adjacent *and* non-adjacent) neighbors. Such context effects contain information that is important for naturalness as well as recognition of speech.

A third, perhaps even more important problem is that knowledge of the language is important for speech recognition. For instance, it is easy to transcribe an utterance if one knows the language. Transcription becomes almost impossible if the utterance is in a strange language, even though the basic sounds remain the same. This is because the listener takes advantage of the contextual properties (phonetics, syntax and semantics) to resolve ambiguities (say due to noise or distortion). We still do not have a very effective way of doing this for speech recognition systems.

2. Historical Survey

2.1 Speech Synthesis

Synthesis of speech cannot be accomplished by cutting and pasting smaller units together. Attention has to be paid to smoothing out the discontinuities in such a process so that the resulting signal approximates natural speech. This is not very simple. The first rule-based systems for speech synthesis attempted to do this. We ourselves were responsible for suggesting a grouping of individual phonemes into equivalence classes, based on their context dependency properties. This led to fewer rules (Rao and Thosar, 1974).

Rule-based systems operate in parametric space – not in the amplitude and time domain – because such smoothing is simpler there. Several attempts have also been made to achieve speech synthesis using speech segments. In order to minimize discontinuities, it would be preferable to have word length or longer segments. This requires storing large repositories of speech segments. To synthesize a given passage, it is also necessary to identify the appropriate segments such that synthesis can be done with the fewest joins. This was difficult earlier but is becoming increasingly possible with the availability and economy of mass storage devices.

Prosody and intonation (modulation of the pitch and volume) are quite important for natural sounding speech. This is a complex aspect and this skill is not easy to acquire even for humans except during their initial years. (This is the reason for accent problems that people have when they learn their second and subsequent languages.) There are in existence speech synthesis systems which replicate the prosodic features of human speech. This involves fairly complex parsing of the input sentences and using rather complex rules to determine the intonation patterns.

2.2 Speech Recognition

The phonetic typewriter of Olson and Belar (1957) at RCA was one of the earliest implementations of speech recognition. It was a hardware system based on a filter bank spectrum analyzer; it incorporated features of state transition networks and the concepts

of non-linear time warping. It is interesting to note that these have once again become prevalent in current systems in recent times.

There were several attempts at speech recognition based on acoustic parameters. Among the early attempts, were use of only steady state information (Raj Reddy, 1967), synthesis based recognition and use of information in transition regions, (Rao, Thosar, 1976; Rao, Paliwal, 1982). Artificial Neural Networks (steady state models), Time Delay Neural Networks as well as Memory Neuron Networks (Poddar Pinaki, Rao, 1992, 1993) were also adopted. Dynamic Programming and Dynamic Time Warping were useful to compensate for speaking rate fluctuations.

In contrast with such purely quantitative approaches, researchers tried to implement intelligent, rule-based speech recognition systems which could mimic the strategies devised by human experts to read and interpret speech spectrograms. In a sense, the Hidden Markov Model (HMM) approach also attempts to (indirectly) capture the speech generation process. It models the dynamics of the human articulatory system as a sequence of static configurations or 'states'. Each state has a range of associated outputs. The temporal relationships between the states are modeled by state transition probabilities. For a given speech signal, speech recognition consists in determining the (most likely) sequence of states that might have generated it. The problem is that to start with, one knows neither the sequence of states nor the outputs (emission probabilities) for each state. Both these are 'hidden'.

HMMs provide an effective and efficient way of representing both the temporal as well as the acoustic variability in the speech signal. Being compatible with (bigram and trigram) grammar models (which we explain later), they can seamlessly incorporate probabilistic properties of language. A limitation is the assumption that parameter value distributions can be effectively modeled in terms of Gaussian and Gaussian Mixture densities (this is not always true). Secondly, they do not take into account the strong cross state dependencies, adjacent as well as non local, that characterize speech. By ignoring the time-wise continuity constraints on proximal states, they restrict acoustic decision making to rather narrow temporal windows. In this sense, they are at a disadvantage compared to the human speech recognition system which takes full advantage of these.

Recent HMM systems use MEL filter based parameter representations; however, they do not exploit knowledge about the articulatory features and constraints of human speech. Human speech recognition is highly robust with respect to time and frequency warping over a very large dynamic range. This is not true for HMMs.

2.3 Grammars

Speech recognition can be robust (tolerant to noise, mispronunciation etc.) if there are constraints. e.g. if the system has to choose between only the four words, North, South, East and West, there is no problem even if the user mispronounces North as "Norse"; the system would surmise correctly that the intended word is 'North'.

More subtle constraints operate when the listener knows the language being spoken. For instance, the sentence, "please turn on the inkjet printer" can be understood even if turn is recognized as churn and jet as set. A sentence wrongly recognized as "Wreck a nice beach" can be corrected as "recognize speech" from higher level context information. Such advantages disappear if the recognition system has no knowledge of the language or the task context.

It is not easy to incorporate such knowledge into a computer. There are however other simpler but reasonably effective techniques. For instance, in any given language, certain word pairs are quite likely and others much less so. (e.g. the combination "good boy" is very likely but not "food boy"). On the basis of such statistics, a system which misrecognises "good boy" as "food boy", could recover and correct it to "good", as intended. Word pair probability data would therefore be useful for improving speech recognition accuracy. This is true even for word triples (trigrams), quadruples (4-grams) and so on. However, simple though this seems, collecting such statistics is a voluminous task. For a vocabulary of n-words, there can be n^2 distinct word pairs, n^3 triples, n^4 quadruples and so on. To compute probabilities, it is necessary to scan a corpus which is an order of magnitude larger in size! It has been reported that, to collect trigram probabilities for English in the 1980s, IBM had a

background job running on all their hi-speed computers in the entire world for several months. Each of these machines spent every spare second that they had on this job. The language corpus they used for this study consisted of the entire set of correspondence in the IBM offices worldwide. Due to such complexity, only trigram grammar models are used in speech recognition. Going beyond this seems to be impractical and cost ineffective.

Speech recognition systems have so far not gone beyond using such fairly simple statistical grammar models. Difficulties of implementation apart, such grammars are not applicable for conversational speech (informal sentence construction, incomplete sentences, unconventional phraseology etc.).

3. Practical Applications

3.1 Dictation Systems

A dictation system needs to recognize not responses to system prompts but freely constructed sentences. There can be no restriction as to the domain of discourse. As against this, it needs to recognize well the speech of only one user. Such systems can be trained to the user's speech; during training the user has to read out special passages provided by the system,

Dictation systems also use tri-gram grammars (which are usually constructed for written speech). My own experience with reputed commercial dictation systems has been that they function quite well when the sentences are well structured, but not so well where sentences are informal as in e-mails.

Several reputed firms have marketed dictation systems for use on regular PCs and laptops. However, this application does not seem to have really taken off. Most users (in India as well as abroad) seem to discard such systems soon after the initial curiosity wears off. One is not aware of any user sticking to such systems for regular use.

3.2 Dialogue Management

We have already emphasized the advantages of speech as a mode of interaction between the human and the computer. How would such a system work? Let us say for instance that one needs information about train timings. Ideally one should need to just ask "when is the next train to Gorakhpur?" To deal with such a question, a computer should:

(a) recognize freely spoken speech (i.e. convert the signal into a sequence of words

(b) interpret that sequence (i.e. determine the intent of the speaker – train departure times, of trains headed for a city named Gorakhpur, whose departure times are later than the present clock time, the earliest among these)

(c) do all this even if some of the words are wrongly recognized

Current technology has not reached that level. Today the user has to give this information in small bits. For instance, a railway information system might give the speaker choice between options available: reservations, train timings, PNR status etc. He has to respond with a single phrase (in this particular case, "train timings"). The next choice might be between trains bound North, South, East or West; the answer could be "North". The next choice would be between Uttar Pradesh, Bihar, Uttaranchal etc. and the response "Uttar Pradesh". Then, between Varanasi, Allahabad, Gorakhpur etc. The user's choice would be "Gorakhpur". In this particular approach, the system tightly binds the speaker to conform to the structure it imposes. He has to respond in terms of one from a set of options it provides at each stage. Such a system is practical, but not very easy to use.

Today, commonly used systems provide somewhat more flexibility; though they still constrain the user significantly. Instead of giving the user a leading role in the interaction, the system retains the lead. It asks him questions and expects him to respond in a predictable way; e.g. to a question "where would you like to go?", the expected answers could be, to Gorakhpur, I would like to go to Gorakhpur, I am going to Gorakhpur etc. The system might fail if he responds with a rather unusual sentence construction, my destination is Gorakhpur or my intention is to end up in Gorakhpur. The system steers the user into a dialogue which it manages so long as the responses conform to its expectation. Dialogue design is therefore a crucial aspect of designing such systems.

'Speak as you like' systems have also been announced. These provide more flexibility but free dialogue of the type that we mentioned at the beginning of this section is still some distance away.

3.3 Transaction Oriented Speech Recognition

Most applications for speech recognition tend to be transaction oriented. The man on the street usually approaches an organization for performing a transaction. Enquiry seeking to elicit specific information (bank account balance, stock price, weather forecast, astrological information), purchase of train, bus, air, cinema or theatre tickets and so on. The conversation that takes place between the user and the service provider therefore tends to be restricted to the specific task on hand (one does not discuss politics or weather with the booking clerk at a railway counter). The sentences used tend to be simple and straight forward. Often, he doesn't even use full sentences. The counter clerk does not necessarily pay attention to every single word in the sentence; much less parse the sentence in a traditional manner. To elicit the intended meaning, in all likelihood, he scans the utterance to determine what the key concept is; that done, he finds out what the relevant key words are. Thus, with minimal effort, he elicits meaning, even when presented only with phrases and ill-constructed or unfinished sentences. Imagine saying to a booking clerk "would you condescend into doing me the favour of booking two tickets in the most luxurious class available, since I, accompanied by my spouse, intend to visit my relations in the township of Gorakhpur, exactly a week from today?" It would be interesting to speculate as to what his reaction would be. Most likely, he will not understand your intent. The point is that such a sentence will not make sense because it is not easy to extract the key concept and the key words from it. On the other hand, the phrase "two first class tickets to Gorakhpur on July, 10th" is adequate, because the key concept (purchase of train tickets) is implied (because it is a railway ticket booking counter); the required key words are all there. The key concept is not implicit if the counter happens to be an enquiry counter which permits a wider variety of key concepts. In such cases, the key concept has to be explicitly stated (eg. arrival time of Rajdhani from Mumbai or platform number of Geetanjali to Bhubhaneshwar; the key concepts are underlined, the others are the key words). Obviously, transactions are much easier to accomplish if the key concept is given out first and then the key words. Reversing this order usually leads to confusion. The same principles seem to apply in most cases of routine day to day conversations between individuals. This approach obviates the need for complex grammars and complicated word processing. (Individuals indulging in routine conversation do not like to expend much effort in background grammar oriented processing.) One often finds that complex sentences which need parsing to be understood (e.g. the boy who stood first in class is going to the cricket field where he scored a century less than a week back) need to be repeated or even rephrased before they can be effectively understood. This is because people in general do not have photographic memories to remember, scan, analyze and extract the meaning from complex sentences. Also, this involves too much effort. Conversational sentences need not be grammatical. They just need to be amenable to easy analysis on the fly. The intended meaning should be extractable without complex processing.

We believe that a key concept and key word based approach would provide a quite effective solution to the problem of natural (spoken) language understanding in a wide variety of situations, particularly in man-machine interaction systems. At the CSR Lab at Tata Infotech, we are actively exploring this approach both for textual and speech interaction with computer systems.

3.4 Speaker Verification -

Speech synthesis and recognition are by far the most talked about applications for speech technology. Lesser known but no less important are several others. There are a number of situations where speaker recognition comes in handy. Being a non-invasive biometric means of identification, it is useful for applications ranging from physical access control for high security areas to control of access to personal computers and information systems as well as attendance monitoring in large enterprises. Unlike finger scan and iris scan systems, use of speech to verify identity is unobtrusive, non-invasive, easy to use and 'friendly'. It is economical (there is no need for expensive - or any - additional hardware) and resistant to variations due to the subject's mood, drowsiness, sore throat etc. Since the parameters it uses cannot be manipulated by the subject, it is quite robust and resistant to imposter attacks. It also does not carry the risk of lost or misplaced cards as in smart-card systems. Verification is possible even at a distance (via phone). It can be

easily blended with other speech based voice response and speaker recognition systems. On the other hand, speaker verification systems are prone to false rejections due to ambient noise, echoes and static. They also require large templates. As of now, they are not very commonly used for high security applications except to achieve higher levels of security in association with other more conventional security systems.

Speaker verification systems have been notably employed for a variety of applications.

- to verify whether a person under 'house arrest' is indeed at home (New York City Department of Corrections)
- to make secure transactions by telephone (Bacob Belgium 1999)
- bank account access via telephone (Allied Irish bank, Buytel's, speaker key technology)
- stock trading via telephone (TimeMac solutions, Australia)
- telephone based account access (Charles Schwab & Co, Nuance, 1999)

We implemented a speaker verification system (at the Cognitive Systems Research Lab. at Tata Infotech) and achieved high performance. In the secret password mode, valid acceptance rate was 98% and valid rejection rate was 97%. False acceptance and false rejection were at 3% and 2%. In the case of public/ common password, the performance was 88%.

3.5 Accent Tool –

With the spectacular increase in BPO enterprise in India, there is an explosive demand for call centre agents with good accent. Staff turnover is very high; so is the cost of manual screening and selection of applicants with good accents. There is therefore a strong case for an automated system that evaluates the accent of applicants for such positions. At the CSR Lab, we also have implemented an accent evaluation system which can be used to automatically screen applicants and very significantly reduce the staff recruitment/ replacement costs in Call Centres.

4. Speech recognition opportunities in India

Several factors work towards making speech recognition oriented applications very attractive in the Indian context.

1. The literacy rates being low, it will be difficult for the majority of Indians to derive benefit from facilities for easy access of information from conventional computer systems; these require literacy and keyboard skills.
2. The Central and State Governments have a clear policy for bringing e-governance to the masses. This can be readily realized only if the masses are able to freely interact with e-governance systems (using speech).
3. The many languages and dialects in India present a challenge as well as an opportunity for speech technology, in providing workable and convenient solutions.

There are significant differences between the phonology of Indian languages and those of the Western World. It is questionable as to how well recognition systems originally designed for western languages will perform if they are merely retuned for Indian languages. Here again there is much scope for developmental work for designing recognition systems specifically for Indian languages.

5. Speech Technology – Outlook for the future

Attempting to predict the future outlook for a modern technology could be very hazardous, particularly in a fast paced area like information technology. Subject to that limitation, let us see what could be the most interesting and desirable developments in this technology in the coming years.

5.1 Recognition based on an Articulatory Model: As we have seen, the HMM based approaches for speech recognition have been reigning supreme for decades now. There has been no real alternative. My own personal preference is for a system that benefits from our knowledge of how the human accomplishes this task. Two important aspects of human recognition deserve emphasis here. Firstly, like all other tasks of human perception, speech recognition too appears to take place in the framework of a generative model – in this case, the human speech production mechanism. Secondly, though the human performs speech recognition 'on the fly', there is strong evidence indicating that he does not do it with blinkers on – one small segment at a time. Rather, he does it on a

more holistic manner, dealing with words, phrases and even whole sentences at a time. HMM does neither of these in a proper sense. The n-gram grammar models do perform a somewhat global optimization. However, phoneme recognition itself takes place in a purely local sense. One might expect that an articulatory model based system more or less of the type indicated below will do well and gain precedence over the HMM based systems before too long.

I strongly believe in taking full advantage of our knowledge of the way in which humans recognize speech. Across modalities, the human perceives stimuli on the basis of the underlying generative model. For instance, to recognize say the hand written word 'and', the human does not deal with the 'picture' alone but unconsciously recognizes it by visualizing the movement of the writer's pen on paper. Thus he can manage even if parts of this word are missing, say because the pen skipped. In general, using a generative model makes the process robust with respect to noise and distortion.

Like most perceptual systems (fig.1), human speech recognition too is strongly based on the generative model. Humans do not recognize speech in a passive open loop manner. They do it as if by visualizing the dynamics of the speaker's articulatory mechanism while producing that sound. He compares the output of the model the speech signal he hears with. He then adjusts its parameters (the position of the tongue hump, whether the lips are rounded, the extent to which the jaw is open and so on) to minimize the difference between the two (fig.2). Once he determines (in this manner) the dynamics of the speaker's vocal tract, he surmises what the speaker has said. All this happens without his being aware of it.

I have for long been enthusiastic about such an (analysis by synthesis) approach for speech recognition by machine (Rao, 1998). The problem however, is that, given the speech signal, it is not necessarily possible to uniquely determine the configuration (e.g. Itakura area functions) of the human vocal tract. It might be practical to do this using a generic, simplified articulatory model, starting with regions where it is easy (e.g. while a vowel is being uttered - initial recognition can be accomplished as a first approximation by using conventional methods). From there, it should be possible to continue tracking the signal by manipulating the articulatory model. Any ambiguities about determining the articulatory configuration at any given point of time can hopefully be resolved by using time and space continuity constraints with respect to how the articulators can move. (If the mathematics suggests two alternative configurations as possible for a given sound, accept only the one that is physically possible for the speaker, given the position of his articulators in the immediate past.) There are reasons to believe that such approaches might yield better speech recognition performance. The international speech research community is increasingly getting interested in such areas. There is increasing interest in statistical modeling techniques that are capable of effectively and parsimoniously characterizing long-span dependency properties in natural human speech (Deng et al, 2003, Seide et al, 2003). It would be desirable to encourage such work locally.

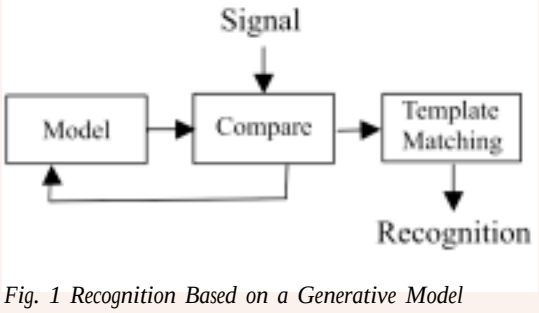

Fig. 1 Recognition Based on a Generative Model

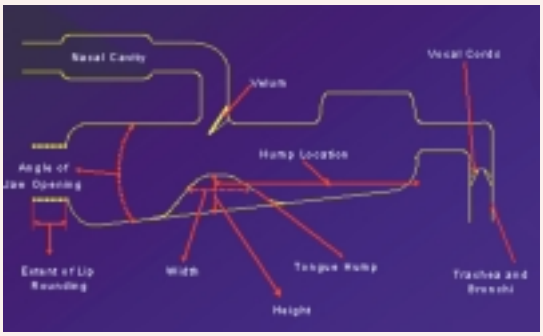

Fig. 2 A Simplified Generative Model of the Vocal Tract

5.2 Holistic Recognition Based on Spectrograms

Looking at the entire spectrogram as a 3-dimensional object is another attractive prospect. Human spectrograph reading experts adopt this method. Treating the MEL filter based spectrogram as a 3-dimensional object and using this for speech recognition also appears attractive. (Segmentation is best avoided because it is itself error prone. It also makes it difficult to take into account the context dependencies and constraints that exist in the speech generation process.) It should be easy to define a metric for the distance between two different spectrograms, in terms of the extent to which one of them needs to be distorted in the 3-dimensions to become the other. In such a system, it would for instance be a simple matter to deal with volume fluctuations and frequency multiplication which the human can handle easily but existing systems cannot.

It is highly desirable for funding research activity along such directions in India.

5.3 Other high priority areas

With the diversity of languages in the country and the need for speech-based man-machine interaction in e-governance, education and other applications, *speech databases in Indian languages* are going to become very important. Compiling such databases would also be very necessary in order to implement speech recognition systems in Indian language and Indian English.

Despite much work and significant progress the quality of synthetic speech appears to be less than good even for corpus based synthesis, particularly in terms of naturalness. Without good, natural sounding speech, man-machine interaction systems will not become very popular. Funding for *high quality speech synthesis systems* would again be very desirable.

Speech based biometric access control and user authentication systems will have a large market worldwide, thanks to concerns about secrecy of personal information and security of transactions. This therefore would be a fruitful area for research.

6. References

[1] H F Olson and H Belar, "Phonetic typewriter", J. Acoust. Soc. Am., Vol. 28, p. 1072, 1957

[2] Raj Reddy, "Computer recognition of connected speech", J. Acoust. Soc. Am., Vol. 42, pp. 329-347, 1967.

[3] P.V.S. Rao and R.B. Thosar, "A programming system for studies in speech synthesis", IEEE Trans. Acoust. Speech and Signal Processing, ASSP-22, pp. 217-225, 1974.

[4] R.B. Thosar and P.V.S. Rao, "An approach towards a synthesis based speech recognition system", IEEE Trans. Acoust. Speech and Signal Processing, ASSP-24, pp. 194-196, 1976.

[5] K.K. Paliwal and P.V.S. Rao, "Synthesis based recognition of continuous speech", J. Acoust. Soc. Am., Vol. 71, pp. 1016-1024, 1982.

[6] P.V.S. Rao, "Is Rule based Acoustic Phonetic speech Recognition a Dead End?", The Second Symposium on Advanced Man-Machine Interface through Spoken Language, Hawaii, USA, 1988.

[7] Poddar Pinaki and P.V.S. Rao, "Hierarchical Ensemble of Neural Networks", Proc. IEEE Conf. on Neural Networks, California, 1992.

[8] Pinaki Poddar and P.V.S Rao, "Temporal Sequence Classification by Memory Neuron Networks", Proc. IEEE-SP Workshop - Neural Networks for Signal Processing III, USA, pp. 181-189, 1993.

[9] PVS Rao, "Are HMMs the last word?", International Workshop on Spoken Language Processing, TIFR, Mumbai, 2003.

[10] F. Seide, J.L. Zhou, and L. Deng, "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM — MAP decoding and evaluation," Proc. IEEE International Conf. Acoust., Speech, and Signal Processing, Hong Kong, Vol.I, pp. 748-751, 2003.

[11] L. Deng, I. Bazzi, and A. Acero, "Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint," Proc. European Conf. on Speech Communication and Technology, Geneva, Switzerland, 2003.