

A. Sen and S. Sanyal

Access to Computerized Information Base by Voice

Proceedings of International Conference on Computers and Devices for  
Communications (CODEC-98), pp 251-254, 1998, Calcutta

Proceedings of the International Conference  
*on*  
**Computers and Devices  
for Communication  
(CODEC-98)**

**Science City,  
Calcutta, India, 1998**

*Editors :*

**S. DHAR**

*Department of Electronic Science  
Calcutta University*

**A.K. DAS GUPTA**

*Institute of Radio Physics and Electronics  
Calcutta University*



**ALLIED PUBLISHERS LIMITED**

---

*New Delhi • Mumbai • Calcutta • Lucknow • Madras  
Nagpure • Bangalore • Hyderabad • Ahmedabad*

# Access to Computerized Information Base by Voice

Aniruddha Sen and Sugata Sanyal

Computer Systems and Communications Group,  
Tata Institute of Fundamental Research,  
Homi Bhabha Road,  
Mumbai — 400 005

## Abstract

With the rapid growth of information networks, the need to access information by dialling is fast increasing. The paper outlines such a scheme, based on Speech Recognition which handles spoken queries and Speech Synthesis which generates spoken replies. The core issue of speech recognition is identification of the underlying linguistic messages and the main problems are filtering out of all sorts of statistical variations and taking care of the differences in phonetic contexts. Dynamic Time Warping and Hidden Markov Modelling are some techniques commonly used. For telephone query, (single) word recognizers, or (multi) word-spotters with somewhat less accuracy, can be used. The main issues of Speech Synthesis are establishing pronunciation for the given text, generating speech which is highly context sensitive and making synthetic speech natural sounding. Several reasonably intelligible synthesizers for Indian languages are already in existence.

## 1 Introduction

We are now watching an explosion in Information Networks. To 'hook into' such networks, a computer is needed. Access to telephones is much wider than access to computers and will continue to remain so in near future. Facility for 'hooking into' such networks by voice will thus make the information accessible to many more.

That such a scheme is feasible was previously proposed by the authors [1]. This paper describes the related methodologies in some details in the Indian context. The query of the user has to be handled by a 'Speech Recognizer' which 'understands' it i.e. deciphers the (textual) message from the input speech signal. This message is then to be presented as a query and when the answer (again, in terms of text) is available, it is passed on to a 'Speech Synthesizer' which

'reads' it out i.e. converts the text into corresponding speech signal.

Both Speech Recognition and Speech Synthesis are open research issues as yet and there are limitations in the systems implemented in India or abroad. This paper outlines some associated research problems, the state-of-the-art in India and indicates how, despite limitations, the proposal can be implemented to good effect. Some future directions are also indicated.

## 2 System Overview

In the proposed system, the user connects to the information base by dialling into the computer which prompts him to speak a keyword from among the given choices. When the user answers, the speech is conveyed to the computer in digitized form. The Speech Recognizer then processes it and passes the textual query it interprets to the query handler which may then access the information asked for from the attached information base and/or may prompt the user to specify his requirements in more detail (i.e. going another level down). The accessed answer/prompt is passed on to a Text-to-Speech synthesis system which generates the corresponding speech in digitized form. The telephone line is connected to the computer through appropriate interfaces. Besides this, Analog-to-digital and Digital-to-analog converters and appropriate cut-off filters are the only additional hardware needed. Inexpensive speech boards, housing the telephone interface, Analog-to-Digital (A/D) and Digital-to-Analog (D/A) converters and filters are commercially available. (See Figure 1).

Some possible information which may be thus accessed are Railway or Airway Reservation status; Weather information or information about Stock Market movements. The only restriction the state-of-the-art imposes is that it should be possible to construct the query with a few simple keywords.

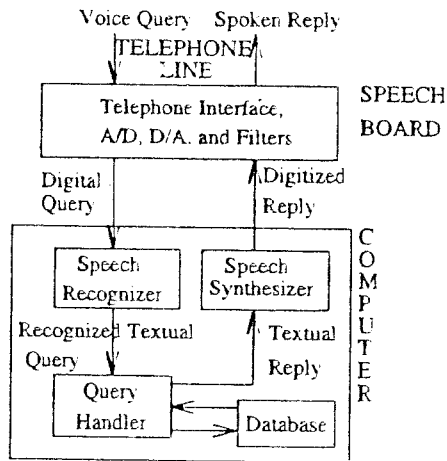


Figure 1: Information access by Voice

Clearly, even if such systems are readily available in international market, it will not exactly suit to our needs. For effectiveness, we need information systems handling queries in Indian languages. Even when the query is in English, the accents of the speakers will be Indian and the accent of the synthetic reply should also be Indian. In other words, the Speech recognizer and synthesizer must be for Indian languages and Indian English.

### 3 Speech Recognition

#### 3.1 Recognition issues

The main issue in Speech Recognition is extraction of the textual information from the signal after filtering out all kinds of variations (e.g. inter speaker variations, speaking rate variations). Another important issue is taking care of the high acoustic-phonetic context-dependency of speech signal. This is especially important when continuous speech recognition is attempted. An additional issue is recognition of noisy speech and telephone speech is noisy in more than one count (e.g. band limitation and line noise).

Normally, speech is represented by a set of features extracted from its short-time spectra. The features chosen for recognition (e.g. Cepstral co-efficients and their differences) are such that their variations correspond to the variations of linguistic information. Speaking rate normalization can be accomplished by techniques like Dynamic Time Warping (DTW) where the features of stored templates and the test utterance are aligned by dynamic programming method with best match criteria [2]. Some appropriate distance measure

(e.g. Itakura Distance) [3] is then applied and the token at the *minimum* distance from the stored template is chosen as the one recognized. Alternatively, in Hidden Markov Modelling (HMM), each token is represented by a state-transition model (Markov chain). The model parameters are estimated from training data. While recognizing, the probability of the model generating the utterance is calculated. The model having the *maximum* probability of generating the test data is chosen as the unit recognized [4].

The token (or unit) of recognition is a whole word in an isolated word recognition system. For continuous speech recognition, it can be either a phoneme or an acoustically coherent sub-word unit determined by special techniques like Vector Quantization. Recognized sub-word tokens are to be combined into words by suitable lexical analysis.

It is well known that in order to attain high recognition efficiency, linguistic constraints (arising from restrictions imposed by grammar, meaning and contexts) are to be utilized over and above the acoustic considerations mentioned above. A suitable Natural Language Processor therefore processes the output of the acoustic recognizer (specified usually as multiple choices with probability values) and constructs the most plausible linguistic hypothesis as the final output.

#### 3.2 Present Recognition Scenario

Internationally, a number of large vocabulary isolated word speech recognizers are now available. However, high degree of reliability has not been achieved so far in continuous speech recognition. Techniques employed are usually HMM or DTW.

In India, a number of medium vocabulary recognizers, based on HMM or DTW technique, have been developed. In the Tata Institute of Fundamental Research (TIFR), a few medium vocabulary (50-200), isolated word recognizers have been developed and at least one of those was utilized for similar query-related tasks. The recognizer is based on HMM technique. Specifically, a strict left-to-right model is employed and continuous Gaussian mixture density is used. The state transitions are restricted to only forward transitions up to two immediately following states. The system is pre-trained by Indian speakers and therefore can accept spoken words from users with Indian accent without any additional training. The recognition accuracy is acceptable (above 90%) [5]. The capacity of the system is currently being enhanced: the vocabulary is being increased to about 200 and the system is being trained by a larger number and wider variety of people (male, female, children). The initial results show

that the recognition accuracy at 4-5 word choice is very close to hundred percent.

This or similar systems can be applied gainfully for the proposed information access systems. However, research on robust speech recognition has not yet advanced far. Therefore, to keep the error rate within limits, it is necessary to keep the vocabulary low. That restricts the range of the system's applications, but in many situations a small vocabulary will still be good enough. Often the choices can be numbered and ten digit-words can thus cater for a wide range of queries.

To speed up information retrieval process, it is however preferable that more than a keyword may be input in each query. Without going to the yet unreliable domain of continuous speech recognition, this can be done by a method termed as 'Word spotting'. Here, the spoken query can be continuous and can contain more than one keyword (e.g. Is there a *First Class* ticket from *Calcutta* to *Delhi*?). The recognizer does not recognize the whole sentence, but scans the length of the utterance for the selected number of keywords. Accuracy offered by this method falls between the accuracies of isolated word and continuous speech recognizers. Research and Development on this type of recognizer is going on in India. The choice between word spotter and isolated word recognizer should be carefully made, depending upon their speed and relative accuracy. If accuracy is less, queries are to be repeated more often and this should be considered while estimating the overall rate of information exchange.

## 4 Speech Synthesis

### 4.1 Synthesis Issues

Speech is highly context-sensitive. Same phoneme (unit of utterance), in the context of different adjacent phonemes, has different acoustic manifestations, with the transition segments carrying vital information for human perception. Generating continuous synthetic speech by 'Cut-and-paste' of phonemes is thereby not possible and this is a major problem in speech synthesis.

This problem is tackled either (i) by taking bigger splices of speech (e.g. syllables) as units and joining them at steady portions of speech, as is done in Concatenation Synthesizer or (ii) by generating speech by suitable production models and varying the model parameters by a set of context-dependent rules which capture the transitions and context variations, as is done in Formant or Articulatory synthesizers. While the former type of synthesizer models the speech production

system in terms of the acoustic-phonetic parameters like energy, pitch and resonance ('Formant') frequencies associated with speech and uses heuristic rules to drive the model, the latter one uses precise mathematical models of movement of articulators, i.e. tongue, lips, jaws in order to generate speech. For more details and comparative studies of different types of synthesizers, Klatt's review paper [6] may be referred to.

Another issue of speech synthesis is text-to-phoneme conversion i.e. generating pronunciation, given the text. In several languages (e.g. English), text and pronunciation do not have a clear correspondence. Text to phoneme conversion is then non-trivial. Applying prosody (accent, in short) is another issue. Capabilities to pronounce correctly with proper accent is very 'humane' and is not easy to be imitated by a machine.

### 4.2 Present Synthesis Scenario

Presently, there are a number of continuous speech synthesizers in India and abroad, applying either concatenation or formant synthesis techniques. Articulatory synthesis is currently at the centre-stage of speech synthesis research and is likely to emerge in a big way in near-future.

In a synthesizer developed at TIFR, Formant Synthesis technique has been applied. In order to generate synthetic utterance, a standard Source-filter model of vocal tract [7] is driven by a comprehensive set of rules relevant to Indian languages. The phoneme repertoire of the synthesizer includes all typically Indian speech sounds e.g. aspirated and retroflexed consonants or nasalized vowels. A few more has also been added to cater for Indian English. The rules, derived after considerable experimentation and inferencing, capture variations of acoustic-phonetic parameters in all possible phonemic contexts. Taking advantage of a very flexible organization of rules which leaves scope for changes without much effort, quality is being constantly improved upon. The synthesizer can generate both male and female voices and can also vary voice types and speaking rates [8].

Several other text-to-speech systems, based on concatenation techniques, have also been developed for Indian languages (e.g. Hindi, Bangla) [9] [10]. The outputs of all of them are quite intelligible. Moreover, it is now possible to generate synthetic speech in Indian languages with acceptable speed on easily affordable machines like Pentium PC. 'Real-time' synthesis is now possible which means that the machine can continuously read text from a file and 'speak' it out, just as a man reads out a book. Such synthesizers can be conveniently applied for 'reading out' messages over

telephone.

However, the outputs of synthesizers in general and the Indian language synthesizers in particular, are not yet too natural-sounding. As prolonged exposure to unnatural-sounding speech can cause discomfort, synthetic speech, therefore, should better be applied to deliver brief messages only. Fortunately, that is what the outputs of most query systems are.

To adapt an Indian-language synthesizer to a specific Indian language, a front-end text analyzer for that specific language is to be developed. Whereas such text analyzers for several phonetic Indian languages (e.g. Hindi, Urdu) have already been developed with a degree of success, the work has just been initiated for Indian English. The successful accomplishment of this task is important. As the language of most of the information bases is English, an Indian English synthesizer will open up many new applications to the users. Reading e-mails over cellular phone is one such use which is rapidly gaining importance.

## 5 Conclusion

Although the scheme can be implemented with acceptable degree of success with the state of the art Indian language speech recognizers and synthesizers, considerable improvement is needed for future.

In order to cater for the need of increased vocabulary, robust speech recognition is to be improved upon. One prospective area is Auditory Modelling which tries to model the non-linear pre-processings our auditory system performs in order to enhance speech and reduce noise. Work has been initiated in this direction. Work has also been initiated in continuous speech recognition and acceptable systems should emerge in a few years.

In speech synthesis, quality and naturalness has to be improved upon for better acceptability. More and more efforts therefore must be directed towards understanding prosody and applying them for speech synthesis. Articulatory synthesis is an area with immense prospect which is not yet explored much in India. It is planned to initiate research in this area in the context of Indian languages in near future.

It may also be added that speech research, in general, is nearing saturation in acoustic domain and the future will observe considerable research in related areas of Natural Language Processing viz. pre-processors for speech synthesizers and post-processors for speech recognizers. Indigenous, language-specific research is therefore gaining increasing importance in the area of speech research.

## References

- [1] Sen, A., Furtado, X.A. and Sanyal, S., 'Technological relevance of speech synthesis in India', *Proc., Comp. Soc. India Convention*, Bangalore, Oct.-Nov., 1996, pp. 38-45.
- [2] Rabiner, L.R. and Schmidt, C.E., 'Application of dynamic time warping to connected digit recognition', *IEEE Trans.*, vol. ASSP-33, no.3, Jun 1995, pp.561-573.
- [3] Itakura, F., 'Minimum prediction residual principle applied to speech recognition', *IEEE Trans.*, vol. ASSP-23, no.1, Feb., 1975, pp. 67-72.
- [4] Rabiner, L.R., 'A tutorial on Hidden Markov Model and selected applications in speech recognition', *Proc., IEEE*, 77(2), 1989, pp.257-286.
- [5] Rao, P.V.S., Bhiksharaj, R., Sen, A. and Mallavadhani, G.R., 'A computer tutor with voice I/O in Hindi', *Knowledge Based Computer Systems Research and Applications*, Eds. K.S.R. Anjaneyulu, M. Sasikumar and S. Ramani, Narosa Publishing House, 1996, pp. 491-502.
- [6] Klatt, D.H., 'Review of text-to-speech conversion for English', *J. Acoust. Soc. Am.*, 67, 1982, pp. 737-793.
- [7] Klatt, D.H., 'Software for a cascade/parallel formant synthesizer', *J. Acoust. Soc. Am.*, 67, 1980, pp. 971-985.
- [8] Furtado, X.A. and Sen, A. 'Synthesis of unlimited Speech in Indian Languages using formant-based rules', *Sadhana*, June, 1996, pp. 345-362.
- [9] Bhaskararao, P., Peri, V.N. and Udpikar, V. 'A text-to-speech system for application by visually handicapped and illiterate', *Proc. of the ICSLP 94*, Tokyo, Japan, 1994, pp. 1239-1241.
- [10] Dan, T.K., Datta, A.K. and Mukherjee, B., 'Speech synthesis using signal concatenation', *J. Acoust. Soc. India*, vol. XVIII(3 & 4), pp. 141-145.