

Aniruddha Sen, Xavier Furtado and Sugata Sanyal

Technological Relevance of Speech Synthesis in India

Proceedings of the 31st Annual Convention of the Computer Society of
India, 30th Oct – 3rd Nov, 1996, pp. 38-45, Bangalore, India
Tata McGraw-Hill pub. Co. Ltd.

COMPUTER SOCIETY OF INDIA



INDIA

THE EMERGING INFORMATION TECHNOLOGY GIANT

*Proceedings of the 31st Annual Convention of the
Computer Society of India
30th October to 3rd November, 1996
Bangalore, India*

Editors

L M PATNAIK
K RAJAN
S SASI KUMAR
SUDHA RAJU
A L RAO



Tata McGraw-Hill Publishing Company Limited

New Delhi

McGraw-Hill Offices

New Delhi New York St Louis San Francisco Auckland Bogotá Guatemala
Hamburg Lisbon London Madrid Mexico Milan Montreal Panama
Paris San Juan São Paulo Singapore Sydney Tokyo Toronto

TECHNOLOGICAL RELEVANCE OF SPEECH SYNTHESIS IN INDIA

ANIRUDDHA SEN*, XAVIER A. FURTADO† and SUGATA SANYAL‡

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

Abstract

This paper discusses speech synthesis and its relevance against the backdrop of the IT revolution in India. Speech synthesis has better short-term prospects than speech recognition and despite limited naturalness, can outperform recorded speech in important application areas like dial-in access to data bases. Indigenous development is imperative for synthesis. Out of the three standard techniques, concatenation and formant syntheses have been tried out for Indian languages with equal success. A real time Hindi formant synthesizer has been implemented on a Personal Computer. For future progress, this highly potent core technology needs industrial support.

1 Introduction

Speech recognition and synthesis, corresponding to speech input and output for computers, are much talked about topics in recent times. Unfortunately, the related works are often viewed with either unrealistic expectations or unwarranted scepticism and a prudent appraisal is rare. This paper attempts to examine several aspects of speech synthesis in an objective manner. Its prospects and limitations against the backdrop of the on-going IT revolution are discussed. Its relevance in India is highlighted and the progress is outlined with the example of a particular synthesizer developed at the Tata Institute of Fundamental Research (TIFR). The paper concludes with a few pointers towards the future.

2 What is Speech Synthesis

Generation of speech by machine, usually a computer, may be defined as 'Speech Synthesis'. A mere reproduction of pre-recorded speech (e.g. reading out a voice mail) is not 'generation' and hence should not be termed as speech synthesis which, in its complete form, implies the capability

*email: asen@tifrvax.tifr.res.in

†deceased

‡email: sanyal@tifrvax.tifr.res.in

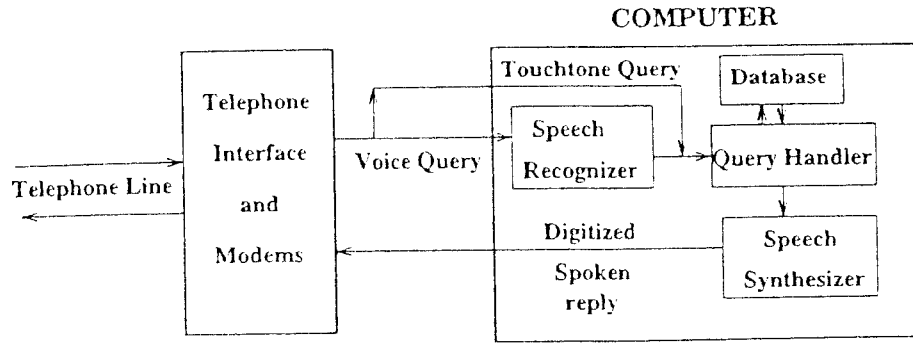


Fig.1 Automatic Answering System over Telephone

for producing speech, given any arbitrary message. Written text is a very convenient means for conveying messages and speech synthesis, for most practical purposes, is text-to-speech conversion.

3 Recognition and synthesis

Usually, recognition and synthesis - speech input and output - go together. However, as the corresponding technology developments are substantially different, there is a strong point for treating them independently. Speech recognition problems are more complex and the progress here is expectedly slower. For example, whether in India or elsewhere, accurate large vocabulary continuous speech recognizers are yet confined by and large to the laboratories whereas unlimited continuous speech synthesizers have hit the market at least a decade ago. Naturally, the latter can have a far wider range of immediate applications. In a 'voice output only' system, graphic interface(GI)-mouse or touchtone buttons (for telephone applications) can be conveniently used for input.

4 Why Speech Synthesis

Making a machine 'talk' has always been an intellectually stimulating exercise in speech and cognitive sciences. Traditionally, synthesizers are known to be useful especially for the illiterate, the blind (for 'reading out' text) and the speech impaired (like illustrious Prof. Hawkins, for 'speaking' through a synthesizer). Recent advances in computer and communication technology, coupled with the explosion in the demand for information, have immensely increased the relevance of speech synthesis for the general population as well. Of particular importance is the possible use of synthetic speech for conveying information of public importance e.g. for announcements at public places like airport, railway station or for 'dial-in' access to computerized information bases (Fig. 1).

5 Synthetic vs. Recorded speech

For the last mentioned application, however, synthesis has a close rival. Conventionally, such information is spoken out by announcers/operators directly or their pre-recorded voices are used. In either case, the choice here is essentially the choice between 'manual' and 'automatic'. When information is extensive and changes very rapidly, the 'manual link' becomes a bottleneck against the ever-increasing demand for throughput. A speech synthesizer offers a solution with its ability to generate spoken messages 'automatically'. When the information is already computerized (which is often the case now-a-days), messages can be synthesized without any human intervention and resulting delay. Even when they are not and are to be keyed in, the operators/announcers may find the task more convenient than straining their vocal cords throughout the long duty hours.

6 Challenges

However, whether in India or abroad, the spread of speech synthesis applications is not as much as one could wish. The basic reason is not only the lack of awareness but also the failure of the state-of-the-art to reach the desired height. This fits into a general pattern of AI endeavours where it is easier to make a machine do an expert's work (like an aeronautical engineering design) than to make it mimic some of the 'common sense' activities like speech, vision, comprehension of language etc. A possible explanation of the 'paradox' is that most of the expert knowledges have evolved over a few centuries at best, whereas 'common sense' was assimilated over tens of thousands of years.

As for speech synthesis, it is now possible to produce reasonably intelligible speech. However, the problems of synthesizing 'natural-sounding' speech have not yet been reasonably solved. For complete text-to-speech conversion, generation of appropriate pronunciation and stress patterns is a prerequisite. This is equivalent to the 'reading' ability in human beings and calls for linguistic and possibly literary knowledge. Not unexpectedly, here also the progress is slow.

In short, whereas some issues in speech synthesis are quite well worked on, some other problems are still open and improvements are likely to come only gradually. Broadly speaking, the work in acoustic domain is fast reaching a point of saturation and in future we are likely to witness slow progress through the domain of language.

State-of-the-art speech synthesizers are quite suitable for supporting applications involving brief output messages like those from most query handlers. However, with the present degree of naturalness, synthetic speech is unlikely to be popular for reading out passages or chapters. Such applications, for some more time, are likely to be restricted to the blind who have no better option.

Which synthesizer

To have any meaningful advantage over recorded speech, the synthesizer should be *unlimited*, i.e. capable of producing any given utterance. Most of such synthesizers are of any of the three broad types, viz. concatenation, formant or articulatory, depending on the approach taken to solve the core synthesis issue of generating speech, given the pronunciation.

In a concatenation synthesizer, speech is produced by combining a sequence of pre-recorded segments of natural speech. As phonemes, the units of speech sound are substantially influenced

by the context, appropriate techniques are to be employed while selecting and concatenating the segments. In formant and articulatory synthesizers, speech is generated by production models. In the former, different types of speech sounds can be produced by varying well-known acoustic-phonetic features like energy, pitch and more importantly, the 'formant' or resonance frequencies of vocal cavities. In an articulatory synthesizer, movements of various articulators like tongue, lip are mathematically modelled and speech sounds are varied by varying the model parameters, thereby emulating the corresponding articulatory movements.

Each methodology has its strengths and weaknesses and with hardly any consensus emerging about their relative gradings, it is safer to say that the quality depends upon not only the methodology chosen but also the actual skill deployed. In general, concatenation synthesizer is the simplest to implement, articulatory synthesizer is the most complex one and formant synthesizer is in between. Simpler the technique, lesser the quality potential but also lesser is the need for specialized skill. A concatenation synthesizer (e.g. LPC synthesizer) is basically corpus-driven, with limited need for specialized skill. A formant synthesizer, on the other hand, is basically rule-driven where the vast body of acoustic-phonetic knowledge available on formant movements is utilized. For an articulatory synthesizer, precise evaluation and optimization of the model are the key issues.

As for speed, simpler the synthesizer higher the speed of execution. However, with modern, fast computers, it is possible to implement acceptably fast synthesizers by employing any of the above techniques and speed is therefore a minor consideration now.

Currently, both concatenation and formant synthesizers are running close in international field whereas articulatory synthesizers are poised to make a big breakthrough. The Indian scenario is briefly overviewed in the next section.

8 Indian Scenario

The last decade witnessed considerable progress in speech synthesis research and development in India. Although we could not yet catch up with the international developments, the gap has been somewhat narrowed down.

As the synthesis issues like text analysis, preparation of speech data corpus, acoustic-phonetic analysis and acoustic-phonetic rule formation (when applicable) are very much language-specific, indigenous R & D is unavoidable. However, a number of areas (e.g. speech signal processing, production modelling) are practically language-independent. There, we need not start from scratch and can utilize the available techniques to the best effect.

As English plays a vital role in IT even in India, prudence demands its inclusion among the 'Indian' languages. However, attempts to use off-the-shelf English synthesizers met with limited success and it shows that in accent and subtleties of pronunciation, English must be 'Indianized' for general acceptance.

Quite a few institutions in India are engaged in speech synthesis research in Indian languages. A fast text-to-speech synthesizer for Hindi/Urdu was developed at Deccan College, Pune [1] and was later on productionized by the Centre for Development of Advanced Computing (CDAC). A talking dictionary-cum-spellchecker in 'Bangla' has recently been announced by the Indian Statistical Institute (ISI), Calcutta [2]. The Indian Institute of Technology (IIT), Madras has worked on a novel scheme where the 'unit' is a 'character' of written text [3]. All three utilized concatenation technique, although the last mentioned one innovatively blended it with formant synthesis. The only reported

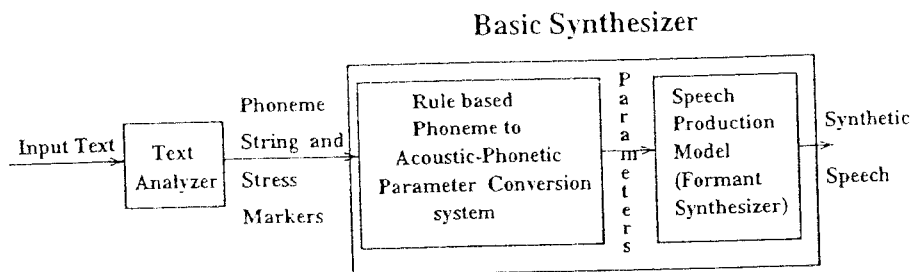


Fig. 2 TIFR Synthesizer Schematic

unlimited continuous speech synthesizer using exclusively formant synthesis technique is from TIFR, Bombay [4]. The Central Electronic Engineering Research Institute (CEERI), New Delhi unit is also engaged in similar research [5].

No noteworthy work has been reported as yet in articulatory synthesis in Indian languages and immediate prospect does not appear to be bright as the necessary ground works in terms of collecting articulatory data or preparing labelled speech data bases have hardly been done. Between concatenation and formant synthesizers, the quality demonstrated so far is nearly as par. It is to be seen whether formant synthesis, with its inherent capacity for improved quality and versatility, can ultimately be the leader.

The following section outlines a typical example of indigenous synthesis effort. It is a summary description of the TIFR formant synthesizer which, from an early research synthesizer stage [4], has been gradually improved upon for supporting possible applications.

9 TIFR Formant Synthesizer

9.1 Functional description

The basic unit (Fig.2) is a phoneme-to-speech synthesizer. It can be connected to any appropriate 'text analyzer' to input phoneme symbols and stress/punctuation markers. Input text can be keyed in or read from a file. Provision exists to by-pass the front-end text analyzer and specify pronunciation directly. Currently, it can accept inputs suitable for generating speech in Hindi or Indian English. The output of the synthesizer is digitized speech for a sampling rate of 16 KHz. Male or female voice, with options for high or low pitch, can be synthesized and the speaking rate can be varied within limits.

The complete synthesizer is a 'software' one. The only special hardware needed is a speech board for playing back the digital speech samples generated. The synthesizer program, written in C, is portable to virtually any platform with minimal modifications. An early PC version was demonstrated in the 1993 exhibition of the Computer Society of India and the latest one runs on a Pentium-based PC in 'real time' which means that one second of speech can be synthesized in less than a second. The significance of breaking the 'real-time barrier' is that with multiple-buffered speech I/O, it is possible to 'read and synthesize' from a text file of any size without any break (Fig.

Currently, a 'Soundblaster' speech board is being used.

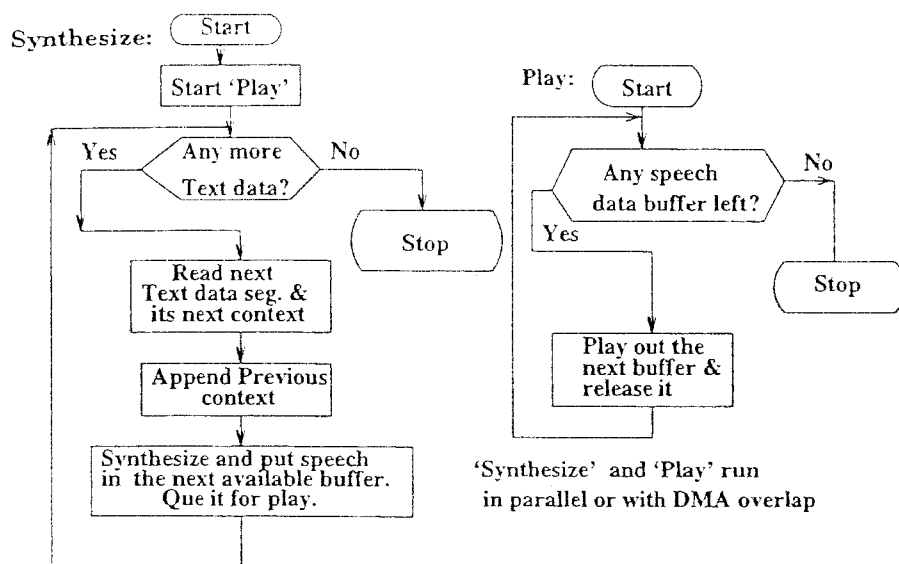


Fig. 3 Uninterrupted 'Read & Synthesize' scheme

By integrating the basic synthesizer with various front-ends, the following demonstration systems have been developed: (i) a Hindi/Urdu text-to-speech system (ii) an Indian English text-to-speech system (text analyzer not yet completed) and (iii) a 'number reader' with Hindi or Indian English options.

9.2 Brief working principle

The synthesizer can be divided into three nearly independent stages, the final one being a speech production model which is essentially a formant synthesizer. It implements a modified version of the standard 'source-filter' model suggested by Klatt [6]. In short, a series of filters, modelling the vocal tract, are excited by some 'source' which is either (i) a train of quasi-periodic pulses, simulating the vibration of the vocal cord or (ii) random noise, simulating the generation of noisy sounds like 's'. This far, the scheme is uncomplicated. But aspects like shaping the source pulses or applying articulatory effects need considerable expertise and determines the quality of synthetic speech. The standard model was expectedly found to be by and large language-independent. Only a few minor modifications were needed to capture a few specific characteristics of Indian speech sounds.

However, the control parameters of this model e.g. energies, pitch, 'formant' frequencies, are to be varied to produce the desired utterances, specified by phoneme strings. TIFR research was focussed on this phoneme-to-parameter conversion issue, which is done by a comprehensive set of acoustic-phonetic rules. The rules are very much language-dependent and therefore cannot be adapted from those reported for English or European languages. The rules are dependent on phonetic context and are therefore complex and extensive. However, they were organized innovatively in a very compact and flexible framework. This allows easy modification by changing the entries in some control tables when the experimental result demands. The rules have been developed for Hindi and Indian English

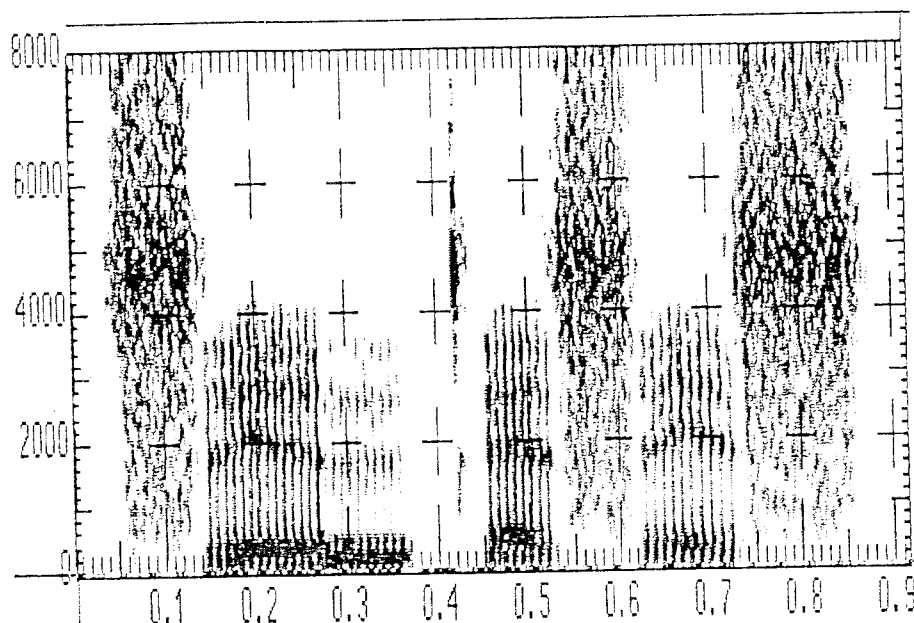


Fig. 4 Spectrogram of utterance 'Synthesis'
as synthesized by TIFR synthesizer

and can be adapted easily for similar Indian languages mainly by changing some table entries.

The front-end of the synthesizer is a text analyzer generating the phoneme string and stress markers corresponding to the given text. Practically any text analyzer can be integrated with the synthesizer by modifying a phoneme look-up table. However, pronunciations atypical to Hindi or Indian English may be approximated to nearby ones. An Indian English text analyzer is currently under development in TIFR.

9.3 Discussion

The synthesizer demonstrated that acceptably fast synthesis in Indian languages is now possible on a PC-platform without any additional hardware like DSP board and without compromising with the quality. Earlier, emphasis was on phoneme-to-speech conversion. As that has been achieved with reasonable degree of success, focus now is on text analysis and naturalness.

10 Conclusions

Speech synthesis in India has progressed enough to support selected applications. Considerably more, however, has to be done before the synthesizers be good enough for general applications. Particular emphasis has to be applied on text analysis.

It must be realized that synthesis is not a 'one shot' problem - here, quality has to be gradually improved to keep pace with the satisfaction of the users. R & D, therefore, is not just an initial

effort, but a continuing activity.

It must be understood that despite current limitations, speech synthesis technology has high potential and can deliver what voice response systems utilizing recorded speech cannot. As off-the-shelf synthesizers with native English accent have limited acceptability in India, indigenous development is unavoidable and has to be supported.

In speech synthesis research and development, various national laboratories are doing reasonably well. However, in India, the interactions between industries and laboratories are feeble and this is detrimental to the interests of both. The industries should realize that speech synthesis is a 'core' technology capable of adding a new dimension to computer industry and catalyzing applications far beyond our current imagination.

11 Acknowledgement

We thank Prof. P. Bhaskararao for permitting us the access of the Hindi text-to-phoneme system for TIFR synthesizer demonstration. We thank Prof. P.V.S. Rao for his encouragement and support.

12 References

References

- [1] 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.
- [2] 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.
- [3] Rajesh Kumar, S.R., Sriram, R. and Yegnanarayana, B. 'A new approach to develop a text-to-speech conversion system for Indian languages', *Proc. of the regional workshop on computer processing of Asian languages*, Bangkok, 1989, pp. 102-109.
- [4] Furtado, X.A. and Sen, A. 'Synthesis of unlimited speech in Indian languages using formant-based rules', *Sadhana*, June, 1996, pp. 345-362.
- [5] Agrawal, S.S. and Stevens, K. 'Towards synthesis of Hindi consonants using KLSYN88', *Proc. of the ICSLP 92*, Alberta, Canada, 1992, pp. 177-180.
- [6] Klatt, D.H. 'Software for a cascade/parallel formant synthesizer', *J. Acoust. Soc. Am.*, 67(3), 1980, pp. 971-995.