INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

IN REPLY PLEASE REFER TO:

H4.SMR. 405/2

SECOND WORKSHOP ON TELEMATICS

6 - 24 November 1989

Introduction and Speech Synthesis

P.V.S. Rao
Tata Institute of Fundamental Research, Bombay, India

<u>These notes are intended for internal distribution only.</u>

LECTURE 1:

INTRODUCTION AND SPEECH SYNTHESIS

FAR, far away in the lonely space, in orbit around a yet-to-be-known planet, an astronaut eagerly looks out the porthole of his space-ship. Suddenly, he spots something that could be a canal or a highway and excited, he starts reporting to headquarters.

A word statesman delivers an impassioned speech at the United Nations. Delegates listen in rapt attention, in a dozen different languages at the same time.

Nearer home, a senior scientist gives detailed instructions to his assistant on the solution of an involved problem. A bank refuses to cash a cheque on suspicion of forgery. A doctor looks carefully at his patient, trying to determine whether his skin rash is merely prickly heat, or something more serious. A secretary takes dictation from a busy executive. A hall full of audience sits spellbound by the music of a famous maestro.

How well can a computer be substituted to serve these situations? A computer can perform a large variety of routine and specialised jobs. But can it replace astronauts, translators, scientists, doctors, bank clerks, stenographers or ;musicians? Can these machines have eyes, ears, and a voice as versatile as their natural counter-parts?

It would be quite simple to connect a television camera, a microphone and a loudspeaker to a computer. But to make the computer extract messages out of signals it receives through these sensory organs and to generate similar signals to convey messages is indeed difficult.

Speech Response by Computers

If a user speaks to such a machine, saying say, the word 'listen', it can receive and retain the corresponding sound wave; even a tape recorder can do that. But to make the computer "recognise" that this sound stands for the word "listen", not for "kitchen" or "kitten" or any other phonetically similar word, is quite another problem. Similarly, if you show it a handwritten English character A, it has to recognise it as such, and differentiate it from a triangle or an H or the line drawing of a hut.

Why should this be so difficult? Why does this difficulty not arise when a computer reads a punched card or a magnetic tape? In such cases, the messages are standardised and generated by machines. Practically no deviation from the specified "norm" is allowed or encountered. Therefore, if you say "listen" (or anything else) in exactly the same tone, with the same loudness and inflexion every time in accordance with a pre-standardised pattern, recognition wouldn't involve much difficulty. The same is true for the visual case.

Man, on the other hand, can follow speech in spite of its varying inflexions and accents. Noise may not bother him too much, even when it is much louder than the speech. At a tea party, for instance, we can follow our friend easily, even when half a dozen others are talking at the top of their voices.

But how does the human brain have so much versatility? Perhaps we shall come to answer this question in our attempts to make machines do similar things. This is another important reason why scientists are trying to make machines that can listen and observe, not merely hear and see.

One might be tempted to believe that this ability involves human like intelligence. We are not aiming at that. For the moment, we are only talking about perception, something very similar to what a stenographer does; substitute one set of symbols (different types of sounds into another (the phonetic alphabet, as of an Indian language). There is no "interpretation" of the "meaning".

Interpretation, too, might ultimately be possible. But a computer sitting at a railway station or at an information desk might be able to perform only a very severely restricted role. It might answer well framed simple questions in a limited sphere of activity. A question like "Who will be the next Prime Minister of India? might elicit a dignified "I beg your pardon". Or it might cause the machine to sulk in silence. In fact, even a human clerk might respond similarly, but that is beside the point.

The anatomy of speech

For the present, let us restrict ourselves to making computers that can speak and listen.

How do we do these ourselves? When we speak, we primarily cause the air molecules to vibrate. It is somewhat after the fashion of a stone dropped into a pond which gives rise to ripples on the surface. For different words, the patterns of vibration are different. The waves are picked up by the eardrum which acts like a microphone. They are then transmitted through the inner ear to the snail shaped "cochlea" where they are translated into electric pulses which travel though the auditory nerves into the brain.

In what mode do the sound waves carry specifications regarding speech? - amplitude, pitch, frequency range, relative phases or something else? This is an important question, because a mechanical computer that listens has to operate on these aspects.

It might take 60,000 decimal digits to describe the details of the sound waves from ten seconds of speech. If however, you wrote this speech as a coded message, you would need only 200 digits for doing it.

This is roughly similar to the problem of describing a man effectively or recognising him from a description. One could do it down to the last wrinkle in his face. Among these descriptions, features like bald head, pointed nose, mole on chin, and brown eyes are useful, but not details like blue shirt, crumpled collar or polished shoes. Features described should be characteristic, unvarying and discernible.

If the information in speech is carried by certain specific elements, scientists thought that it should be easy to extract these agents; what remains should naturally be unintelligible gibberish. They, therefore, tried to "filter" away several things from speech waves, trying to see if intelligibility suffers. But do what you will, distort it, chop off several frequencies or mix up their relative phases, speech remains reasonably intelligible. Obviously, none of these aspects carry speech information, exclusively.

Fig.

Model of a speech spectrogram. Horizontal axis represents time
(0 to 0.4 sec), vertical axis shows frequency (0 to 35000 cycles)
and the heights of the ridges correspond to amplitudes.

Perhaps we shall get a clue if we observe the way speech is generated and recognised by man. Our vocal chords, driven by a stream of air from the lungs, vibrate to produce sound. The frequency of this sound can vary within limits. The sound generated, however, is not a pure tone, and has overtones at integral multiples of the pitch frequency. Their relative amplitudes depend on the way we keep our lips, tongue, jaw and so forth. It is the difference in their relative strength (or energy) that distinguishes an "ah" from an "ee or "oo".

Fig.
Diagram of the generation of human speech

If you say "ou" as in "ouch" the pattern starts like that for "aa" and gradually changes to that for "oo", closely following your mouth movements. The plosive bursts of sound when you say "p" or "k" and the hiss when you say "f" or "s", all have different spectral characteristics. In these cases, since the vocal chords do nót vibrate, the sounds are called unvoiced; there is no longer a distinct fundamental frequency or overtones.

Speech spectrogram

It takes a three dimensional surface to represent the variation of speech energy at various frequencies with time. Such a shape is called a speech spectrogram, and can be represented in two dimensions also, using intensity as the third dimension. Looking at such spectrograms, one finds three or four major ridges in the pattern, where most of the energy is concentrated. These are called "formants" and their pattern roughly represents the shapes, sizes and "activity" of the oral, nasal and pharyngeal cavities.

Since these are closely connected to the movement of the articulating organs, it is possible to distinguish different speech sounds merely from their spectrograms.

During hearing, the nerve impulses carried from the inner ear also have a rough correspondence with the spectrogram. In fact, deaf persons have been trained to recognise speech in this form, indicating that all the necessary aspects of the information in the speech are preserved in this.

Assume a computer capable of extracting the speech spectrogram out of a speech wave. It would then store standard reference patterns for different speech sounds (called phonemes). Every time it hears a sound, it would compare it with the reference patterns and recognise the sound.

Even this is not easy. For different speakers, patterns for the same sounds vary as widely as hand writings. Even for the same speaker, mood, stress and intonation cause wide variations. Worse still, patterns for each phoneme are very much influenced by its neighbours. This is similar to the way individual letters get modified by their neighbours as the hand writes, smoothly going from one letter to the other, without discontinuties. In speech, this effect is far more pronounced, and for our purposes, far more distressing. It is impossible even to demarcate the boundaries between adjacent phonemes; they merge so thoroughly.

This problem has its effects in generating speech by machine. If we try to form synthetic speech by cutting a length of tape containing natural speech into individual phoneme length strips and then reassemble them into arbitrary phrases, the result will be even less natural than a handwritten sentence cut up into individual letters and put together to make new sentences. It will be hardly intelligible. Even if one starts with very standard speech sounds; the result will be like trying to generate handwritten messages using a special typewriter. The ear is very sensitive to discontinuties, and smoothing is very important.

It is obvious that speech analysis and recognition are closely connected problems. To generate a spoken sentence, formant patterns for each phoneme are obtained, strung together, and smoothed out. The spectrograms are then generated, and the final wave form computed. Since all this is mathematical manipulation, it is very easy for a computer to do it. This is thus converted to a voltage signal, which is fed to the loudspeaker. Recognisable sentences have already been generated. Even this speech would sound a little monotonous and mechanical, in the absence of stress (change in pitch) and accent (emphasis on certain syllables).

In an approach developed at the Tata Institute of Fundamental Research, the phoneme are divided into equivalence classes and subclasses to facilitate formulation and application of context dependency rules. These rules are explicitly provided for as separate subroutines. Synthesis, even to the level of the final acoustic signal, is accomplished in the computer.

It turns out that the classification of phonemes into equivalence classes on the basis of their context dependency properties is closely parallel to the classification on the basis of articualtory properties.

Very roughly speaking, speech recognition is the reverse process. As already mentioned, the problem is made more difficult due to the effects of "merged neighbours". The steady state patterns for each phoneme may not be reached even momentarily. Any speech recognition system should therefore incorporate an ability to account for the distortions introduced by "neighbours".

It is clear that speech recognition by machine is feasible. But is it possible that a computer can do this as well as we can? This is perhaps an unfair demand. We have a large store for pertinent information, our memory. We adapt ourselves and learn to follow an accent or a particular mode of speech. We can use linguistic information and the context to help us follow speech, if there is any ambiguity. It may be possible to incorporate only some of these features into a machine.

While one should not be too optimistic, it is possible during our lifetime that the voice answering your questions about spaceship reservations or schedules might be that of a computer.

TALKING WITHOUT A VOICE

Fig.

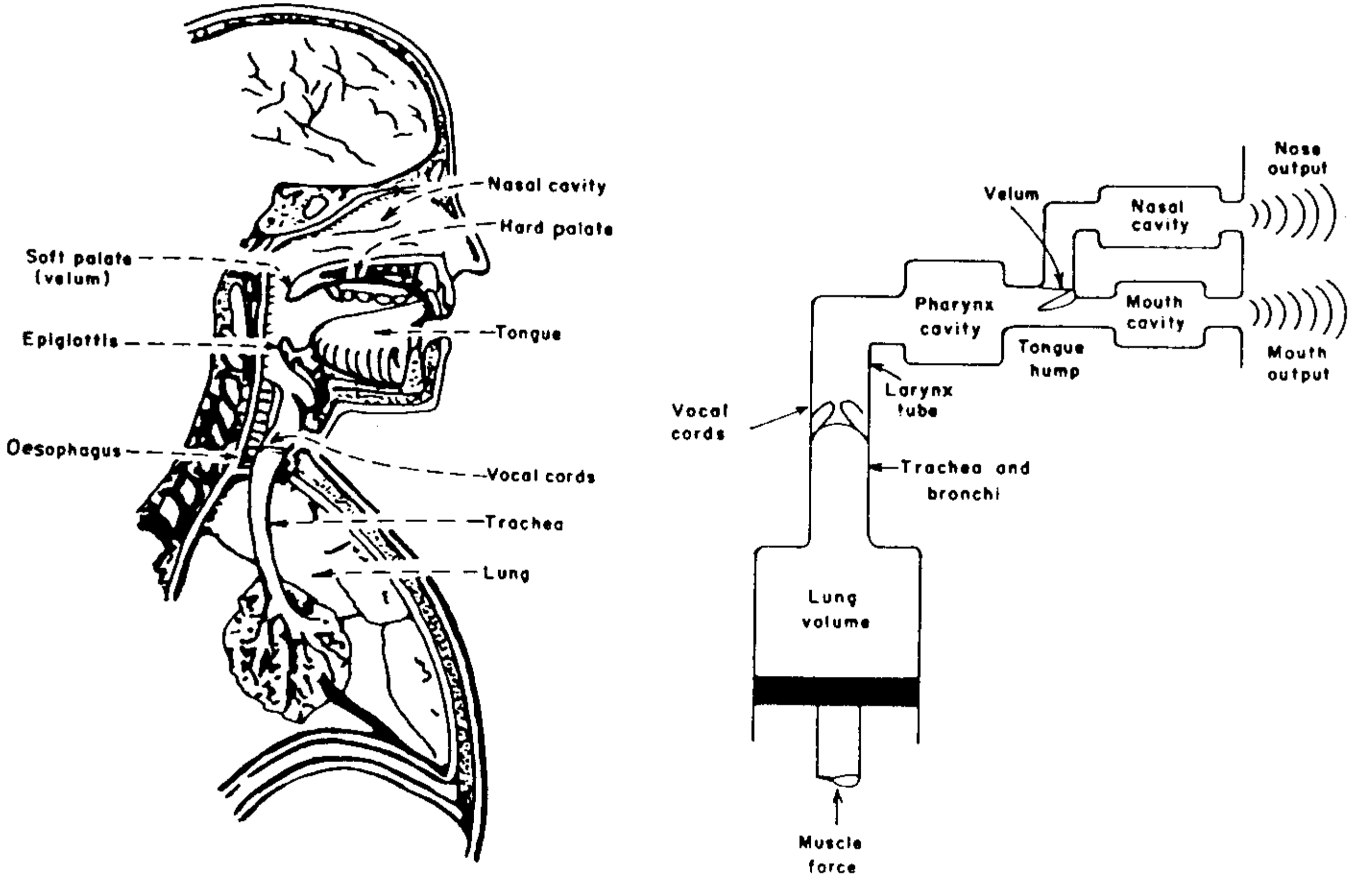
Accurate synthetic speech can now be produced with the aid of a controllable, computer generated model of the vocal tract developed at Bell Telephone laboratories, USA. The method, stored in a computer, is actually a geometric description of vocal tract displayed on an oscilloscope and, at the same time, hear the sound which corresponds to the displayed shape. By flicking switches and turning knobs at a computer console, the researcher can change the shape and sound simultaneously.

In order to synthesize whole words or syllables realistically, transitions are needed between basic sounds. Shapes corresponding to basic sounds are defined by the researcher at the console. The computer then can automatically interpolate sequences of transitional shapes between one basic shape and another. These sequences correspond to the motions of the human vocal tract when full words are produced.

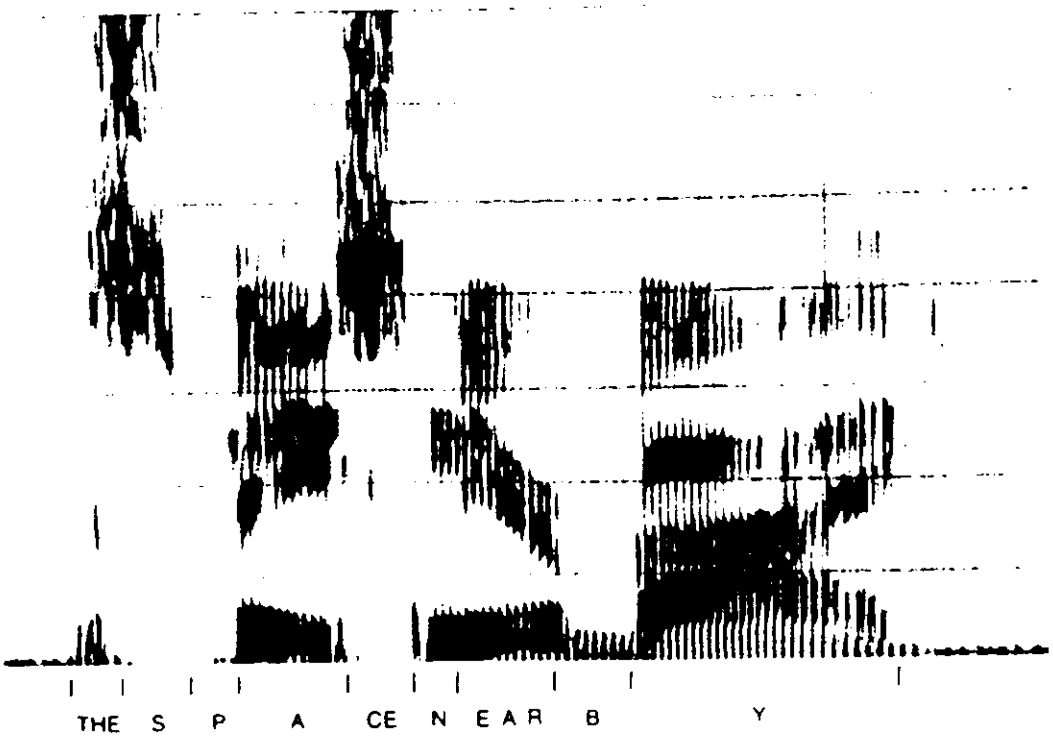
This information may be useful in devising a more efficient means of encoding and transmitting speech signals.

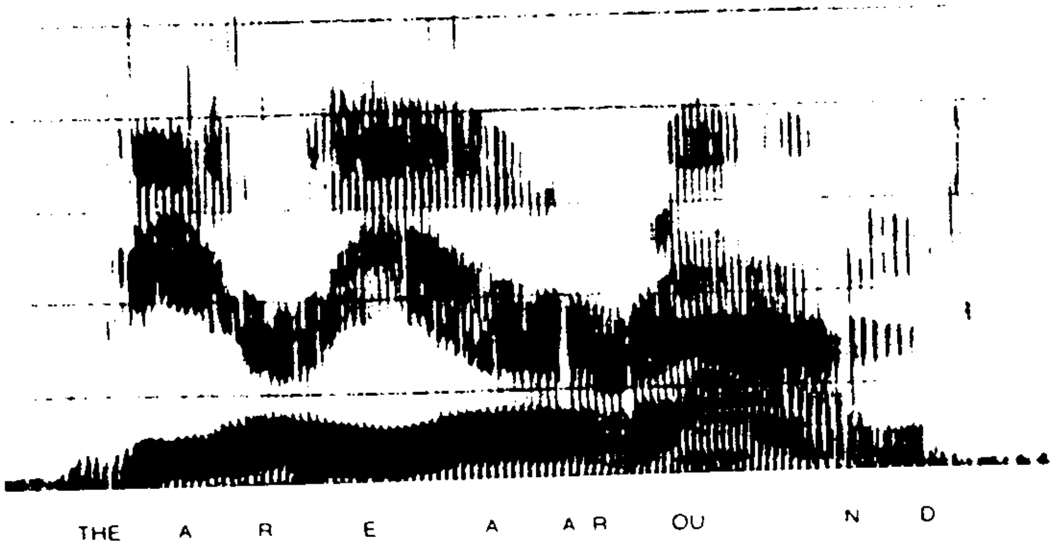
A computer generated oscilloscope display of the vocal tract showing changes in the position of tongue, lips and pharynx.

PVSR/ygk
16.10.89



Speech production system and its schematic representation.


THE S P A CE N E A R B Y


THE A R E A A R OU N D