COMPUTERS THAT TALK - NEW DEVELOPMENTS AND APPLICATIONS

Dr. Russell P. Smith
Managing Director Pulse Data International Ltd

Dr Smith is the Managing Director of Pulse Data International Ltd, a Christchurch-based electronics company. Pulse Data designs and manufactures text and computer access systems for visually handicapped people and also data communications products. Its New Zealand activities include the distribution of high-tech medical instruments from leading overseas manufacturers.

Russell was born in Christchurch in 1944. His secondary education was at Mana College, Porirua. He received a Bachelor of Engineering (Electrical) with Honours from Canterbury University In 1967 and completed it PhD in 1972. His Doctoral research was in wide bandwidth underwater sonar systems.

After a short period lecturing in Electrical Engineering at Canterbury, he joined Wormald Vigilant Ltd to lead a R&D team in the design of the Sonicguide, an ultrasonic sonar guidance system for the blind.

In 197O he became the founding manager of Wormald International Sensory Aids Ltd. a company set up to market the Sonicguide internationally and to develop progressively a comprehensive range of instruments for the visually handicapped people.

In 1988 following a management buy out involving eight of the senior managers, the company became Pulse Data International Limited and Russell became Managing Director.

Under Russell's guidance Pulse Data has become recognised as a leading supplier of innovative adaptive systems for the visually handicapped. The company has subsidiaries in Australia and the USA, and distributors in a dozen other countries.

Pulse Data's most outstanding export sales success has been with talking computers for totally blind people.

Introduction

The evolution of machines that talk has been extremely slow and only in the last two decades has there been Impressive progress. The advent and proliferation of the, digital microcomputer and the development of new signal processing techniques have combined to make the latest generation of speech synthesizers more than adequate in meeting application requirements.

This paper reviews the evolution of synthetic speech and describes a modern generation product.

Early Attempts to make Inanimate Objects Speak

Because, the human species is the only one endowed with speech, the ability to communicate in this way has always been highly prized. Not surprisingly then, over the centuries, man has tried to imitate the complex sounds of speech for advantage, ethical or otherwise.

Most religions ascribe speech to their gods, since clearly gods must be superior to man so they cannot be mute. Zealous priests in early times frequently tried to make their idols "speak" directly to the people as a way of giving the message greater impact. In the early Christian era when many idols were torn down, some were found to contain tubes to carry the voice of the priest from a remote position. The Head of Orpheus, a famous Greek oracle at Lesbos, and statues at Alexandria broken down in the 4th century, had complex voice tubes built into them for this purpose.

Effective though these deceptions were, the ancient world made little progress in synthesising speech. It was not until the 18th century that real advances were made in creation recognizable speech sounds from machines.

Christian Kratzenstein in Russia in 1779 won the annual prize in a contest of the Imperial Academy of St Petersburg for constructing an instrument to produce the human vowel sounds using organ-like resonators.

His solution involved a set of oddly shaped cavities excited by reeds like those of a mouth organ. Fig. 1 shows the shapes of the resonators for each of the live vowels. Apparently the imitation of human vowels was reasonable but no one, including Kratzenstein understood flow they functioned.

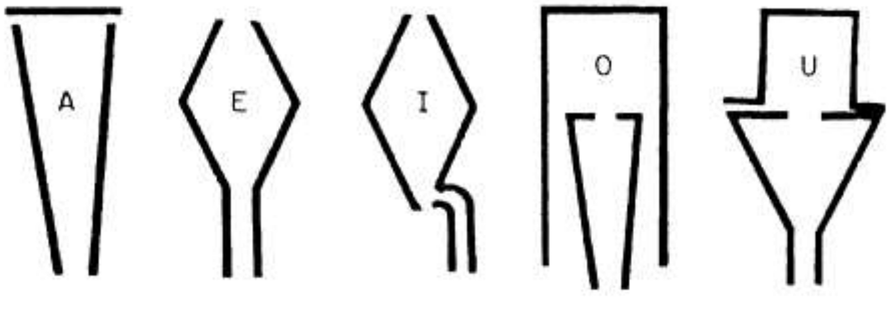


Figure 1: Kratzenstein's resonators for synthesis of vowel sounds. The resonators are actuated by blowing through a free vibrating reed into the lower end. The I sound is produced simply by blowing into the lower pipe without a reed. - After Paget (1930)

More significant progress was made by Wolfgang von Kempelen in Vienna in the 1780's. He used drone reeds from a bagpipe to excite a series of pipes tuned to different frequencies in the hope of creating the vowel sounds. Instead he discovered that they all produced the same vowel sound, just at a different pitch. This lead him to postulate that different speech sounds involve resonances of different frequencies within the vocal tract, and that pitch, the fundamental frequency of the vocal cords, is not involved in defining the individual speech sounds. These different resonant frequencies are now known as "formants".

To test his theory, he tuned all his pipes to the same pitch and introduced obstructions into the pipes to create additional resonances. After some years of experimentation lie was able to obtain most of the vowels and several consonants.
The next obstacle he faced was that he couldn't combine the sounds to produce syllables and words. The concatenation of the individual sounds through keyed excitation of the individual pipes produced harsh unrecognisable sounds. This was also a significant discovery, since it showed that in human speech, each individual sound must merge into the previous and next sounds and may need to he substantially modified to make this transition. This phenomena known as "co-articulation" has been the focus of much research in the last few decades.

Not easily put off by disappointments, von Kempelen concluded that smooth transitions from sound to sound would occur only if all the sounds were generated through a single "mouth", as occurs in nature. After several more years of painstaking experimentation, in 1791 he arrived at his talking machine, a device which models virtually every aspect of human speech production. Fig. 2 shows the construction of the machine.

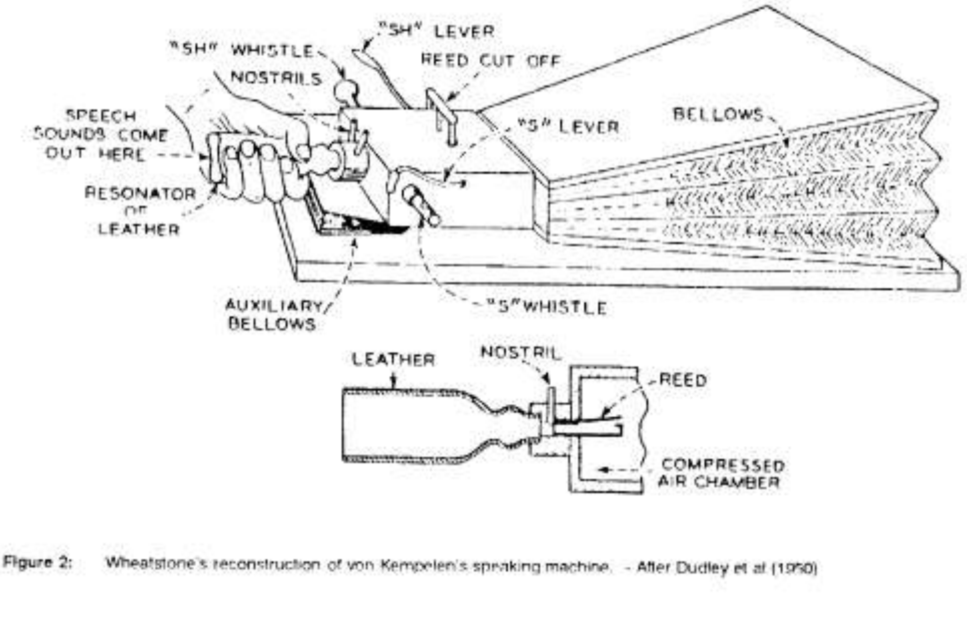


Figure 2: Wheatstone's reconstruction of von Kempelen's speaking machine. - After Dudley et al (1950)

In the hands of a skilled operator the machine could speak whole phrases in French and Italian. By pressing on the main bellows with in elbow, air is forced past the reed causing it to vibrate in a monotone. A lever on top of the "wind trunk" can cause a rod to disturb the airflow over the reed to produce in sound. Other levers allow air to escape through special pipes to give the "s" and “sh" sounds, which are known as "fricatives".

The vowel sounds and “l”, “w” and "y" sounds are made by sealing off the nostrils and manipulating the “mouth” shape to produce the formant resonances. The small bellows adds compliance to the system which is needed to generate the plosive sounds such as "h", "p".

Although some sounds could not be reproduced by substituting the nearest available sound, von Kempelen could deceive his audiences and produce a realistic simulation of speech.

In the 1930's Sir Richard Paget in England made a major contribution to the understanding of the frequency content of vowels and sonic consonants. He had such cut incredible ear for music that he taught himself to identify individual formant frequencies in speech and recorded them. The table in Fig. 3 shows the frequency bands for vowels which he estimated and superimposed (the circles), in measurements made with modern instruments. By teaching himself to control the formant of his own vocal tract, he deduced correctly that at least two formants are needed to give a vowel sound and it is the interval between the formants that defines a specific vowel.

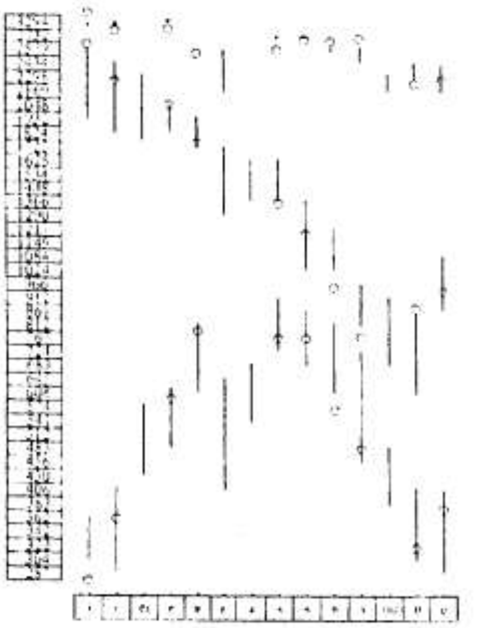


Figure 3: Pagets's vowel resonance chart. The vertical bars indicate the ranges of resonance for the various vowels. The circles are typical formant frequencies measured by modern instrumental methods. - After Linggard (1985)

He also concluded that although there is no harmonic relationship among the formants and the fundamental frequency (pitch) in speech, a good singer can carefully control the vocal tract so that all resonances are in fact harmonically related. The modern generation of “rap” singers have obviously not discovered this essential ingredient of singing quality.

The Nature of Speech

The process of articulationof human speech can be described with reference to the cross-sectional diagram of the human head in Fig. 4

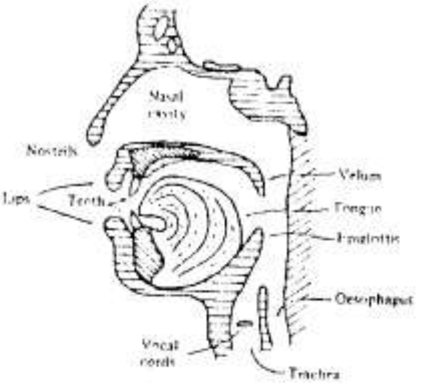


Figure 4: Cross-section through the human head, showing the speech articulars. After Linggard (1985)

The vocal tract is a non-uniform tube approximately 17 cm in length running from the vocal cords to the lips. The cross sectional area at different points along the tube can vary from zero (lips closed) to about 20 square cm (mouth open, tongue low), as different sounds arc produced.

The nasal tract which is brought into play for the "m". "n" and "ng" sounds can he switched in or out by movement of the velum. The nasal tract is a fixed cavity about 12 cm long and about 60 cubic cm volume.

Sound can be generated within the vocal system in three ways:

i) By forcing air through the vocal cords causing them to vibrate in a similar way to the lips of a musician playing I brass instrument. ("Voiced" sounds such as the vowels are derived from this source.)

ii) By creating a constriction at one of several possible positions in the tract to cause severe turbulence in the airflow. ("Fricative" sounds such as "s" and "sh" are produced in this way.)

iii) By completely closing the tract at some point, building up pressure then abruptly opening it, (“Plosive " sounds such as "b" and "d" are created by this process.)

All three types of sound wave have a wide spectrum. The vocal cord oscillation gives a pulse train with individual pulses being I to 4 msec in width, with fairly sharp transitions giving significant energy out to 5 kHz. The fundamental frequency (pitch) can range from 50 - 500 Hz for adults and higher for children. The fricative sounds are noise like with energy over the band from 3 – 8 kHz. The plosives comprise brief bursts of turbulence noise leading into voiced sounds and so have elements of' both these components.

Fig 5 shows the typical waveform of' the pulses at the glottis (vocal cord opening) and the resulting waveform at the lips. Fig. 6 shows a waveform of the word "woosh".

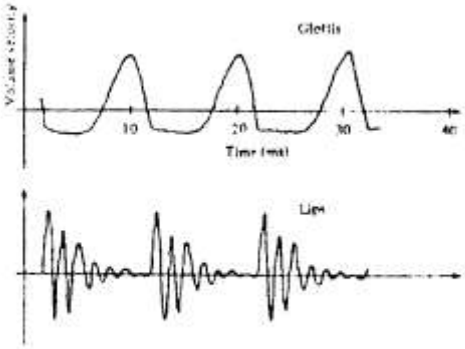


Figure 5: Typical time waveform of volume-velocity pulses at the glottis and at the lips. After Linggard (1985)

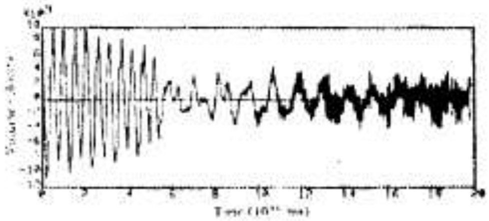


Figure 6: Time waveform of typical voiced and unvoiced speech. The word spoken is "woosh". After Linggard (1985)

Apart from showing the obvious periodicity of the voiced sounds and the noise-like nature of the fricative "sh" the time waveform is not very helpful in understanding the nature of different speech sounds. Spectrograms - plots of the energy in narrow spectral hands as a function of time are much more informative.

Fig. 7 shows an example of a typical spectrogram. This is for the word "seat." and it clearly shows formant frequency paths of the vowel and the wide hand contribution of the fricative "s" and plosive “t.” (In a spectrogram, "blackness" is an indicator of sound energy in a particular filter band at a particular time.)

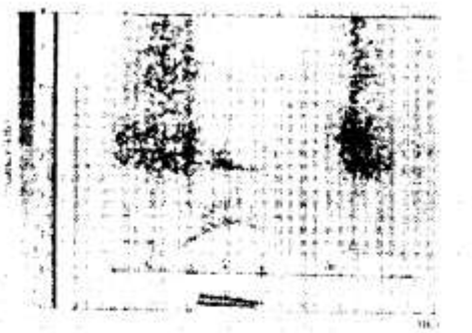


Figure 7: Spectrogram of 'seat', narrow band. Dotted lines show the trajectories of the four formants of the /i/ vowel. After Linggard (1985)

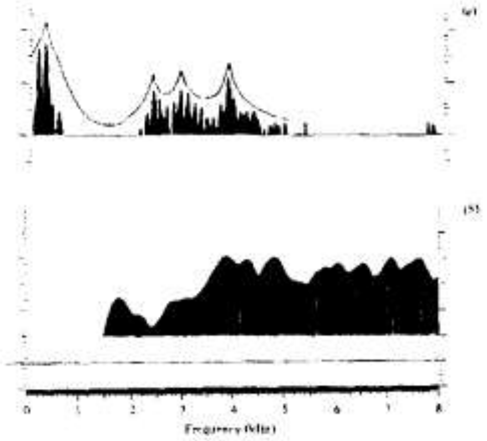


Figure 8: (a) Cross-section of spectrogram of Figure 7 from vowel portion of 'seat'.
(b) Cross-section of spectrogram of Figure 7 from fricative in 'seat'. After Linggard (1985)

Fig. 8 shows vertical cross-sections through the spectrogram during the midpoints of the vowel and fricative portions. The peaks in the envelope of the vowel occur at the four formant frequencies.

This form of spectral analysis suggests it should be possible to generate tile "ea" vowel sound (represented in phonetic symbols as "/i/") by feeding a pulsed waveform resembling the glottal pulse train (Fig. 5) into a filter with poles at the lour formant frequencies and appropriate gain values. This approach does in fact work, and all the vowel sounds can be created accurately in this way by choosing different formant frequencies.

Fig. 9 shows the first three formant frequencies of the vowel sounds in English. It is unnecessary to include more than three formants in order to get clearly recognizable vowels.

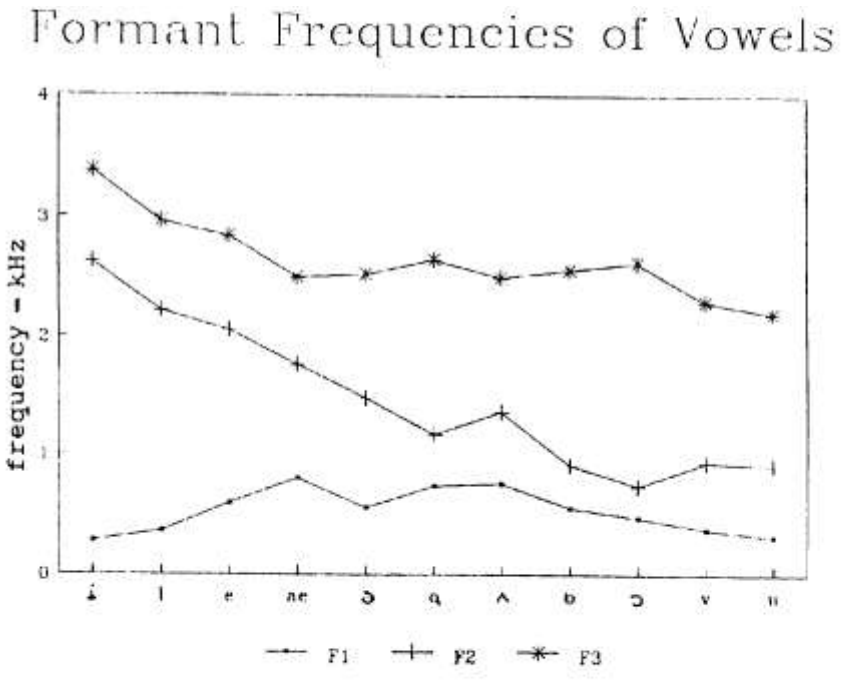


Figure 9: Variation of the first three formant frequencies of the English vowel sounds.

So synthesis should be easy! Unfortunately as von Kempelen out found two centuries ago, the need for smooth transitions from one sound to another makes the concatenation of speech sounds very difficult. This process is perhaps the single most difficult problem facing designers of synthesizers. When it is considered that the formants are produced by the continuous movement of the jaw, tongue and lips during voicing it is perhaps not surprising that these transitions will never be abrupt in human speech generation.

The curved trajectories of' the "ea" vowel in "seat" are needed to achieve a smooth flow from "s" to "ea" to "t". The trajectories of this vowel in "beat", "eat, "ease", etc, are all quite different.

Fig. 10 is a spectrogram which shows tile variation of the first three formant frequencies throughout a complete sentence.

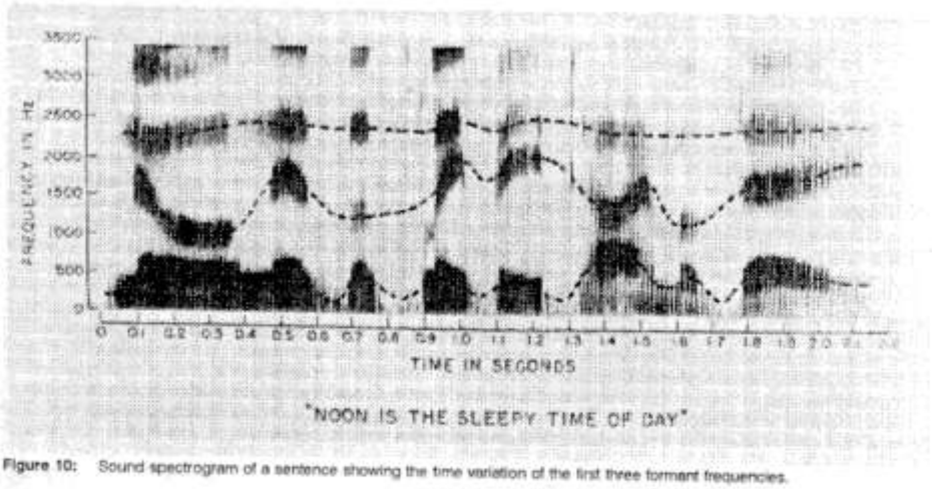


Figure 10: Sound spectrogram of a sentence showing the time variation of the first three formant frequencies.

Evolution of **Electrical Models of Speech Production**

One of the first known electrical speech synthesizers was the VODER (Voice Operation Demonstrator) developed by Homer Dudley at Bell Telephone Labs in 1939.

The VODER, shown diagrammatically in Fig. 11, comprised ten band pass filters spanning the frequency range from 0 to 7500Hz and fed by either of two sound sources - a relaxation oscillator giving an approximation of the glottal waveform with pitch controlled by a foot pedal; and a random noise to generate unvoiced (fricative) sounds. A wrist bar enabled the operator to select between the two sources.

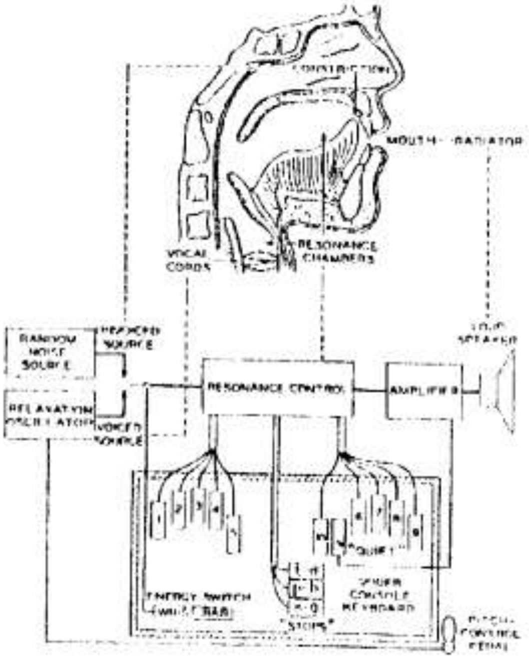
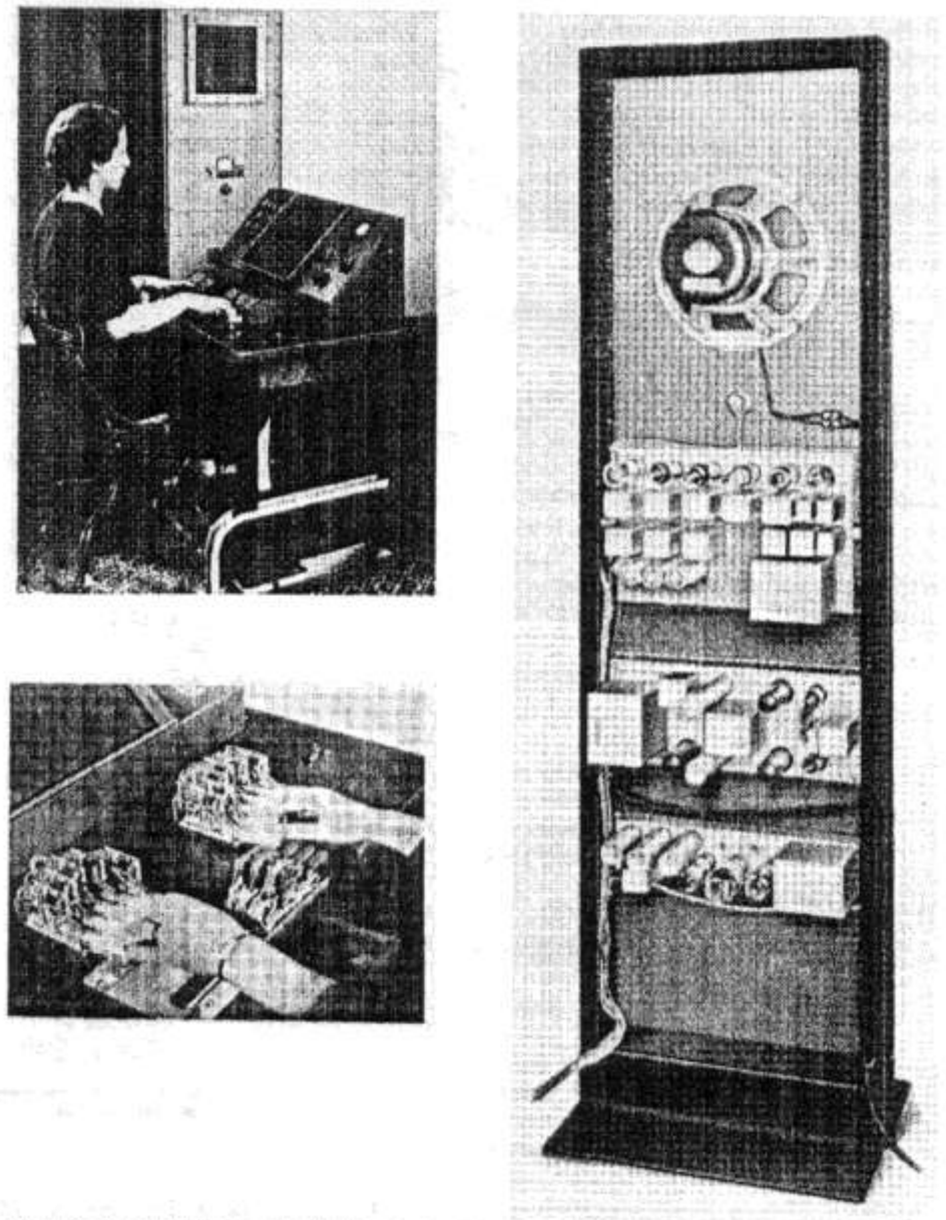


Figure 11: Schematic diagram of the electrical speaking machine, Voder. After Dudley et al (1939)

The operator had a potentiometer under each finger to control the output of˙ each filter. To give the fast attack needed for the plosive sounds, three further controls called ”stops” were included. The model was a close analog of the human vocal system and produced intelligible if unusual sounding speech.
(Audio Demonstration 1 - Refer to Klatt 7987)

Its main drawback was the high skill level needed to operate the controls to achieve the correct sounds and smooth co-articulation. Training took several hours per day for a year or more.

The VODER was a hit at the 1939 World Fair in New York and its trained operators were able to "play" speech to order. Figs. 12. 13 and 14 show the actual machine.



Figures 12, 13 & 14: Photographs of the VODER synthesizer. After Dudley et al (1939)

Impressive though the VODER was, the use of fixed filters has since been abandoned due to the most difficult part of the process, that of producing, the formant variations, being almost impossible to specify or implement.

The Haskins Laboratory's 1951 "Pattern Playback" solved the problem of creating and storing formant frequency patterns by allowing actual spectrograms to be painted onto an optically scanned belt.

Fig. 15 shows the Haskins system diagrammatically. The tone wheel allows a set of harmonics of the pitch frequency (120 Hz) to he distributed across the frequency axis of the transparent spectrogram belt. The density of the palmed lines then determines the amplitude of the formant frequency components as detected by photodiodes. Audio demonstration 2 (refer to Matt 1987) gives an example of the speech which results.

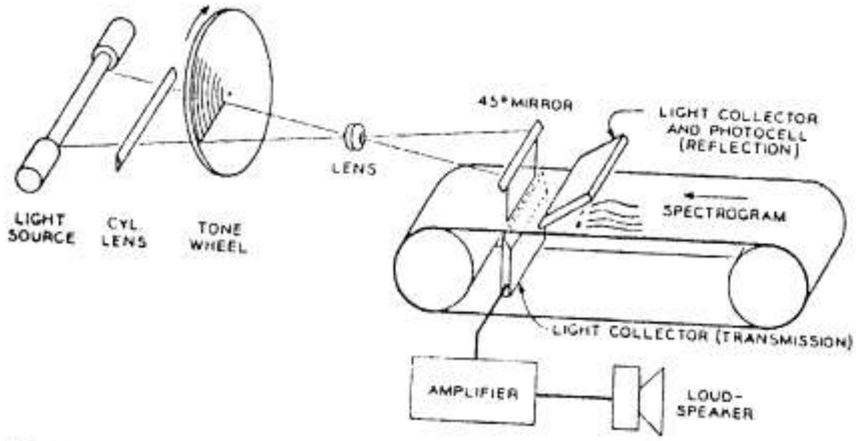


Figure 15: The Haskins Pattern Playback. After Cooper et al (1951)

Fig. 16 shows a later version of the same idea.

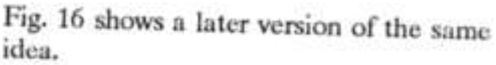


Fig. 16 shows a later version of the same idea.

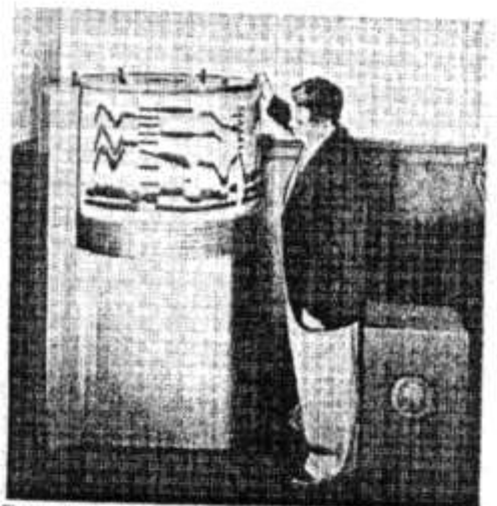


Figure 16: Speech synthesis from shadow patterns of the formants. After Schott (1948)

The VODER and the Pattern Playback were early examples of formant synthesizers. In other words they synthesized speech by blending together appropriately changing formant frequency signals to simulate speech. Another class of model known as articulatory synthesizers directly model the actual vocal tract and its articulators (tongue, lips, jaw, velum).

The DAVO (Dynamic Analogue of VOice) shown diagrammatically in Fig. 17 is typical of articulatory synthesis models. The, vocal tract is represented by a cascaded series of short cylindrical tubes, each section being approximated by an electrical transmission line analog.

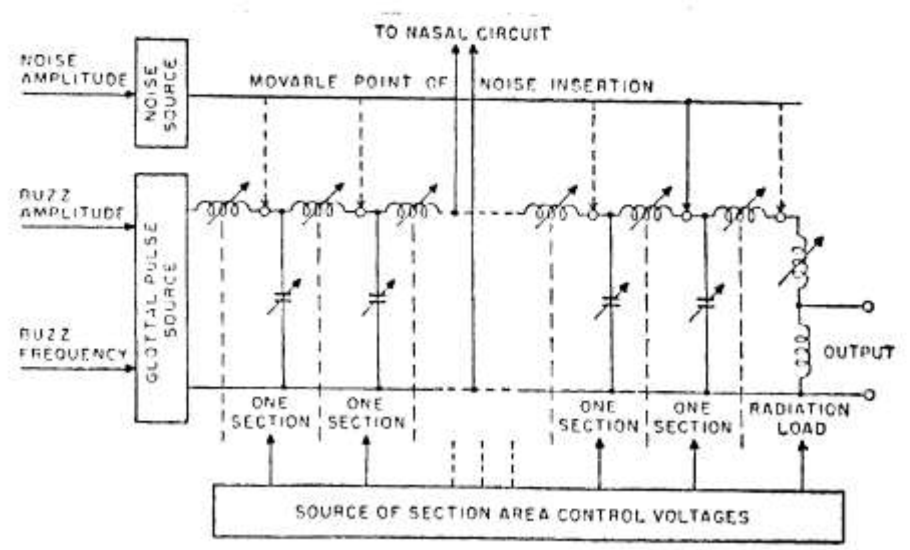


Figure 17: The DAVO (Dynamic Analog of the VOcal tract) synthesizer, consisting of a ladder network of inductors and capacitors Each section models a short section of the vocal tract or nasal passage. After Rosen (1956)

The sound sources comprise the usual pulse train at pitch frequency and a wide band noise source for fricatives. The noise can be injected at various points to model the different points in the vocal tract where turbulence can be created for the different fricatives.

The cross-sectional area of the tract at any point, as modelled by the capacitative and inductive elements of the associated transmission line section, can be changed dynamically by recorded control voltages. A nasal circuit with a fixed transmission line model can be switched in as needed.

This sort of model is potentially the most accurate because every element of the human vocal system is modelled. There is also no problem with co-articulation because the model operates in real time and can only change "shape" at the same rate as tile human vocal tract. Smooth transactions are inevitable.

Audio Demonstration 3 (refer to Klatt 1947) demonstrates the DAVO's performance capabilities.

Unfortunately, determining the time-varying area functions required to produce accurate speech is very difficult so the full potential of articulatory model has not yet been realized.
The most important modelling technique to emerge in the last 20 years or so has been Linear Predictive Coding (LPC). Rather than directly attempting to model the vocal system or assemble sounds from formant frequency components, LPC attempts to model the speech signal on the basis of its predictability.

Linear Predictive theory states that a sampled waveform) from a system with limited degrees of freedom, can be predicted at any instant from a linear weighted sum of its past values.

Fig. 18 shows an ITC synthesizer in a simplified form. The "Linear Predictor" is effectively a segmented delay line and weighted summing circuit in which the weightings are provided its coefficients periodically.

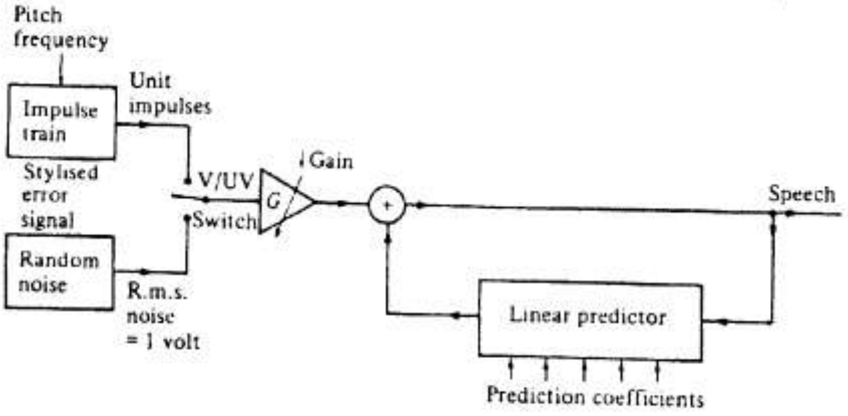


Figure 18: Practical LPC synthesiser. The excitation e(n) is replaced by stylised sources: unit r.m.s. random noise, or a unit impulse train at pitch frequency. Amplitude is controlled by G, gain of amplifier. After Linggard (1985)

The linear predictor can be considered to be a time variant filter which models the vocal tract. Because the vocal tract shape changes only slowly so the filter coefficients need to be updated only comparatively slowly. (Typically every 20 msecs).

Also, since the vocal tract can be modelled as a series of resonators at the formant frequencies, the LPC model call be an all pole filter. In practice 10 or 12 poles are sufficient to accurately model 5 formant frequencies.

To obtain the Predictor coefficients it is possible to analyse a section of real speech, then calculate the coefficients needed to minimize the difference between the real speech and the synthesized version. These will be the optimum coefficients to recreate the original speech.

Audio Demonstration 4 (refer to Atal et al (1971)) gives examples of a sentence resynthesized in this way with different numbers of poles in the LPC filter. It shows that 12 poles is sufficient to create speech which is virtually indistinguishable from the original.
Speech synthesis by all three methods described above is nowadays carried out using microcomputers to model the processes digitally, rather than attempting to use real circuit elements.

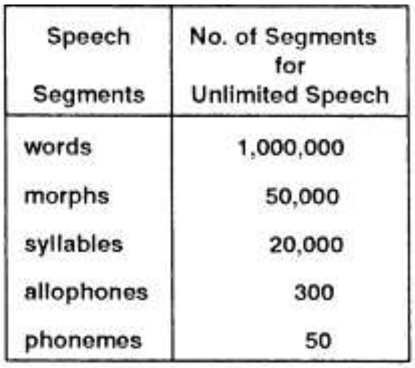
All three types of vocal tract model described are capable of producing speech which is very close to human speech, provided the input data is derived from sections of real speech. The more significant problem is to derive the parametric data which can enable any arbitrary sentence to be spoken. Success with synthesis of arbitrary text has been much more limited because of the difficulties of calculating the required coefficients.

This is the field that my company. Pulse Data International Ltd has been working in for the last 4 years.

In choosing a method for generating speech of unlimited vocabulary electronically, one has to find a compromise among a number of conflicting factors:

Speech segment choice
Vs
No of segments required for vocabulary Speech production method
Vs
Data Storage efficiency Speech quality
Vs Vocabulary size

Figure 19 shows the number of segments required to create unlimited vocabulary for various segment types.



Speech Segments	No. of Segments for Unlimited Speech
words	1,000,000
morphs	50,000
syllables	20,000
allophones	300
phonemes	50

Figure 19: The number of segments required to create unlimited vocabulary for various segment types.

It is clear that if an unlimited vocabulary is required, small speech segments must he concatenated.

Fig. 20 shows the data storage requirements and hit rates for speech generated by different synthesis or recording techniques. Digital recording by PCM produce tape recorder quality storage but at the expense of a data rate of 100k bits/sec and storage requirements of 64K bits per word.

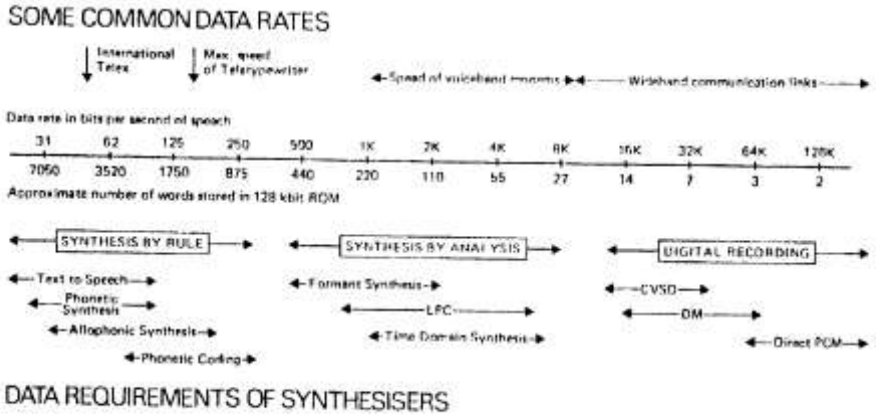


Figure 20: The data rates associated with various speech synthesis methods. After Bristow (1984)

Synthesis by rule from elementary speech segments, such as phonemes or allophones can he achieved with data rates and storage needs 3000 times reduced. In this diagram, speech quality increases steadily from left to right.

Applications requiring high quality speech of very limited vocabulary will be well served by PCM or similar techniques. For unlimited speech, however, quality must be compromised to allow synthesis by rule from small segments.

Conversion of Text-to-Speech

Synthesizers of unlimited vocabulary must be able to take coded text in say ASCII form and convert it to speech regardless of the particular letter strings involved. This process is known as text- to -speech conversion and is fundamental to all synthesizers of' this type

The conversion process is shown diagrammatically in Fly. 21.

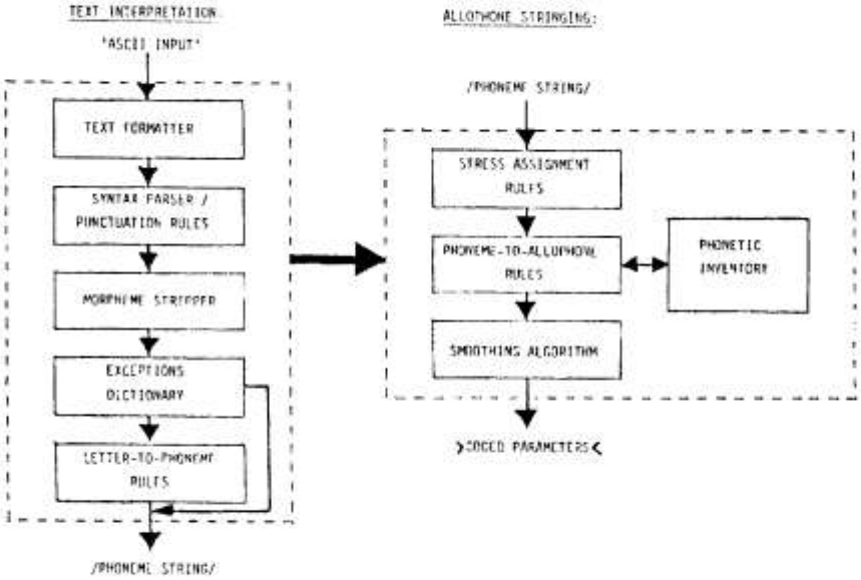


Figure 21: Functional components of a Text-to-Speech system. After Dills (1984)

Text is first formatted by the interpretation, for example, of "Mr." to "Mister". "Rd" to "Road", "1,234" to "One thousand two hundred and thirty four". etc.

It then goes to a syntax parser (not included in small systems) for words like "read" which may for example he pronounced "red" or "read" according to context. Punctuation pausing is also introduced at this point.

Next a morpheme stripper breaks down words such as "helpless" into "help-less". This reduces the number of exceptions and rules needed in the following sections. Translation to a phoneme string is then accomplished from stored sequences from the dictionary if the pronunciation is unusual or by hundreds of rules which define the phoneme sequences for every conceivable letter sequence.

Next the stress pattern within each word and pitch variations within phrases or sentences is applied in accordance with stress assignment rules.

A phoneme to allophone translation then selects variants of each phoneme according to neighbouring phonemes and stress patterns.

Finally the interpolation process ensures allophone transitions are smooth before coded parameters are calculated for driving the synthesizer.

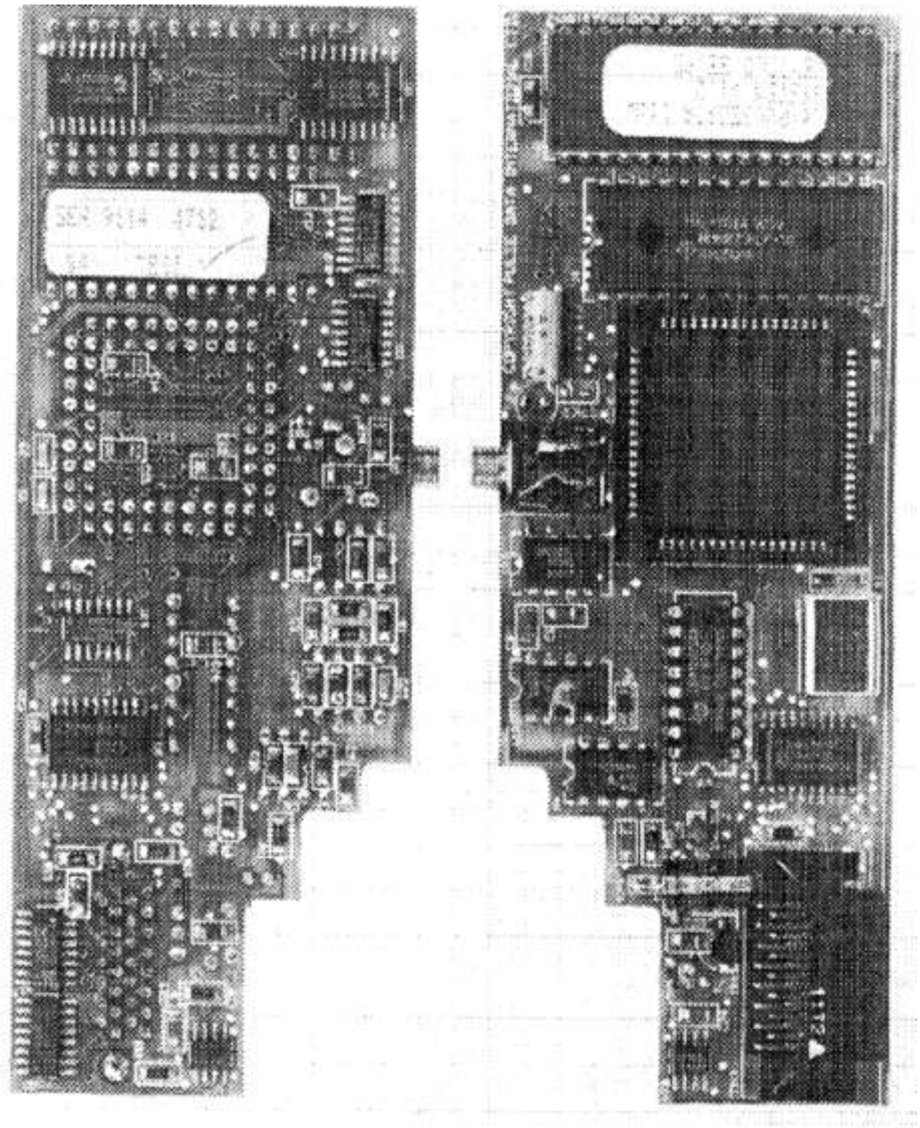
Real Synthesis Products

The first commercial synthesizer with unlimited text to speech was produced by Votrax (a division of the Federal Screw Works in 1978. It used a simple formant synthesizer and low pass filters on a chip and was programmed to generate 64 individual phonemes from simple letter-to-phoneme rules. Audio demonstration 5 (refer to Klatt (1987)) gives a simple of its speech.

In the 12 years since the Votrax chip was introduced substantial progress has been made in programmes to translate I mm text to elementary speech sounds.

Keynote GOLD a state-of-the art synthesizer for talking computers, recently develop by Pulse Data International Ltd in New Zealand, is a good example of the improvements achieved.

Photographs of the synthesizer, packaged to fit inside a laptop computer are shown in Fig. 22.



igure 22: Keynote GOLD synthesizer designed to fit into laptop computers. (Scale: Full size)

The synthesizer chip is a custom mask-programmed 12-pole LPC device. This chip contains the complete 12-pole LPC filter and digital signal sources representing the glottal pulse train it pitch frequency and the random fricative noise. It also includes a Digital-to-Analogue converter to produce in analogue speech signal from the LPC filter.

The text-to-speech conversion is accomplished by a 10 MHz 8088 microprocessor running a programme of almost 250K in length. This enables accurate pronunciation of all but the most unusual words. The text to speech programme is based on conversion algorithms developed by Berkeley Speech Technologies in California. It converts from text to allophones and the LPC filter converts these into speech.

Limitations of Existing Products

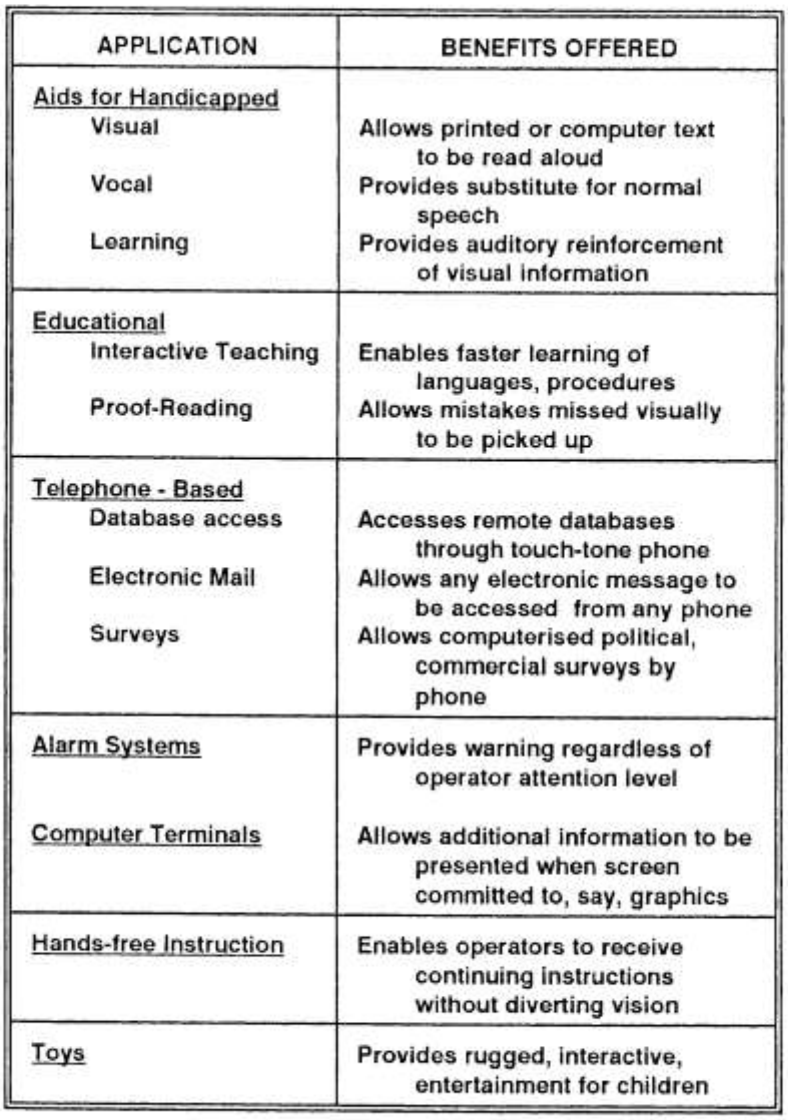
Prosody is the term which describes the variations in pitch, stress and pausing that human beings apply to speech to reduce ambiguities in interpretation and add interest to the sound. Surprisingly, this is one of the most difficult features to create with a text-to-speech synthesizer. The determination of which points to modulate the voice and where to pause come easily to a human reader, but defy simple rules.

Another limitation on unlimited text-to-speech systems is naturalness. Synthesizers like the. Keynote GOLD can achieve intelligibility almost as good as real speech, but they sound like robots. The process of building words and sentences from small speech segments seems to generate a harshness and stilted articulation that is immediately recognized as synthetic.

The most significant progress likely in the next decade in unlimited text-to-speech systems will be in overcoming these two weaknesses.

Applications

Since speech is the most commonly used form of human communication and computers are now almost unavoidable in every aspect of life, it is surprising that talking computers are not commonplace. Speech synthesizers seem to some extent to be solutions looking for a problem. The most significant applications are at present rather specialized as shown in Fig. 23.



APPLICATION	BENEFITS OFFERED
Aids for Handicapped	
Visual	Allows printed or computer text to be read aloud
Vocal	Provides substitute for normal speech
Learning	Provides auditory reinforcement of visual information
Educational	
Interactive Teaching	Enables faster learning of languages, procedures
Proof-Reading	Allows mistakes missed visually to be picked up
Telephone - Based	
Database access	Accesses remote databases through touch-tone phone
Electronic Mail	Allows any electronic message to be accessed from any phone
Surveys	Allows computerised political, commercial surveys by phone
Alarm Systems	Provides warning regardless of operator attention level
Computer Terminals	Allows additional information to be presented when screen committed to, say, graphics
Hands-free Instruction	Enables operators to receive continuing instructions without diverting vision
Toys	Provides rugged, interactive, entertainment for children

Figure 23: Applications of unlimited vocabulary Text-to-Speech systems

Conclusions

Speech synthesis has now developed to the point where text of unlimited vocabulary can be converted to highly intelligible, it somewhat unnatural sounding speech.

For limited vocabulary needs, synthesizers can store very natural-sounding speech very efficiently (compared with PCM).

Synthesizers can be made small enough to fit inside the smallest laptop computers and look set to be an important computer display device for the future.