# The Analysis and Synthesis of the Singing Voice

by

Ken Lomax

Oxford University
St. Hugh's College
Trinity 1997

DOMI NUS ILLU MINA TIO MEA

Thesis submitted for the degree of Doctor of Philosophy
at the University of Oxford

# Acknowledgements

I would like to thank my supervisor Ian Page for giving me the wonderful opportunity of studying for a DPhil at Oxford. My thanks too to my co-supervisor Dr John Coleman who has provided invaluable guidance and draft-checking throughout my research and allowed me to take up residence in the Phonetics Laboratory. Andrew Slater, the technical support officer has given unstinting and first class technical support on the many occasions I have clogged the machines.

I am most grateful to my examiners Professor Burt Rosner and Dr. Christophe d'Alessandro for their comments and suggestions. My thanks too to Xavier Rodet at IRCAM who directed me early in my research and to Perry Cook at CCRMA and Johan Sundberg at KTH who allowed me to visit their research institutes. These trips were possible thanks to generous support from St Hugh's College, the University Chest, the Computing Department and the Phonetics Laboratory.

My thanks also to Julia, Barbara, Tim and Jo for keeping my social life intact, and lastly and most importantly, my thanks to John and Nancy for nurturing my interest in music, offering support and giving me an optimistic outlook, so essential to completing my DPhil.

## Contents

# Table of figures

# Chapter 1

## A review of speech and singing synthesis

### Contents

# Introduction

For millennia dawn choruses have accompanied the morning sunrises. These daily performances serve as a reminder to mankind that we are not alone in our ability to sing. However, our capacity to combine words, emotion and music in song is unique. With its infinite subtleties, the singing voice is a means through which emotion may be expressed, be it joy, sorrow, anger or serenity. At least this is true for the competent singer. When a child first learns to sing, the subtleties of emotional expression are of secondary importance to the ability to sing a melody in tune and with the correct rhythm. When a child does so, we do not say that the performance was unacceptable due to a lack of emotion but that it shows promise, with the belief that emotion and feeling will follow.

Having worked for a major music synthesizer company for two years, I have noticed that although much research and development has been carried out to improve the synthesis of musical instruments, the synthesis of the singing voice has remained out of bounds. The reason for this is at least twofold. Firstly, with its infinite variations in timbre, expression and articulation, the task of synthesizing the singing voice is considerably more complex than that of synthesizing musical instruments. Secondly, although, for example, a synthesized trumpet will convince most listeners of its authenticity, it will not convince an experienced trumpetist. In a similar vein, we are all experts in the human voice, and unless the quality of the synthesized voice is exceptionally high, the illusion will not succeed.

However, if singing synthesis were possible, the potential would be truly enormous. It might, for example, be possible to synthesize particular singers singing new songs. In this case, it would no longer be necessary that great singers from the past are confined to recordings from the past. We could for example have Maria Callas or Nat King Cole 'sing' new songs. Secondly, if it were possible to synthesize these voices in real time, it might be feasible to control the synthesizer with one's own voice and thus have the illusion of singing a song with the voice of a great artist.

People who are confronted with these possibilities fall into two groups. The first agree that the potential is bewildering and exciting. The second believe fervently that it will never be possible to synthesize a realistic voice. Their principle justification is that emotion cannot be simulated and that the synthesized voice will therefore be forever confined to a robotic quality. Of course, I do not agree with this view. More importantly I do not agree with its premise. It is after all not necessary to synthesize emotion itself, only to synthesize the effect which emotion has on a voice. There is no reason to suggest that this is not possible.

In this thesis I describe a means through which singing synthesis may proceed, and I describe the construction of a singing synthesizer. There are two goals towards which I have aimed:

> to synthesize the voices of great singers from the past singing new songs and
>
> to transform the voice of a novice singer into that of a great artist in real time.

Some progress has been made towards both objectives. The approach is somewhat multi-disciplinary, including aspects of computing, music, physics and phonetics.

This thesis contains seven chapters. The first provides an overview of past research in speech and singing synthesis. The second to fifth describe the development and implementation of a software singing synthesizer. The sixth chapter describes the design of a related application termed a singing impersonator. The seventh and final chapter draws some general conclusions.

Audio examples cited throughout the thesis and listed in Appendix A may be heard both on the accompanying audio tape and on the Web page

> http://ouplsun.phon.ox.ac.uk/~lomax/DPhilAudioEgs.html.

Reports of artificial speech production can be traced back to at least the seventeenth century with Don Antonio's talking statue [Cervantes, 1950]. Spectators who thronged to see this reported miracle were not aware of a concealed tube running from the statue's mouth to a man who from his secluded position would speak to the amassing crowds. Myth takes us further back to approximately 1500 BC and the singing Head of Orpheus. When young, Orpheus sang with a voice of such beauty that not only the animals but trees and shrubs would follow him on his walks. Following a misunderstanding with the women of Thrace in which his arms, legs and head were torn off and thrown into the sea, his head was found on the shores of Lesbos. This was taken and placed within a stone wall from where his voice could be heard miraculously singing on [Ackermann *et al.*, 1981].

Only from the mid-eighteenth century have other more reputable approaches been taken in the pursuit of speech and singing synthesis. A thorough review is given in both [Klatt, 1987a][1] and [Flanagan, 1972]. For an introduction to the singing apparatus [Sundberg, 1994b], [Sundberg, 1987], [Seashore, 1938] and [Miller, 1986] may be consulted.

This chapter addresses three areas in speech and singing synthesis, each of which is incorporated in my singing synthesizer:

i. low-level synthesis models,<br>
ii. high-level synthesis controllers and<br>
iii. singing synthesis controllers.

This is preceded by an introduction to speech and to the natural process of speech production.

[1] The audio examples cited in [Klatt, 1987a] may be heard on the Web page http://www.icsi.berkeley.edu:80/eecs225d/klatt.html.

## Speech and the natural process of speech production

The human vocal tract spans approximately fifteen and seventeen centimetres in a female and male adult respectively. In this are housed the vocal cords, mouth and nasal cavity (Figure 1.1). Over many millennia humans have learned to control this system to communicate through speech and language.

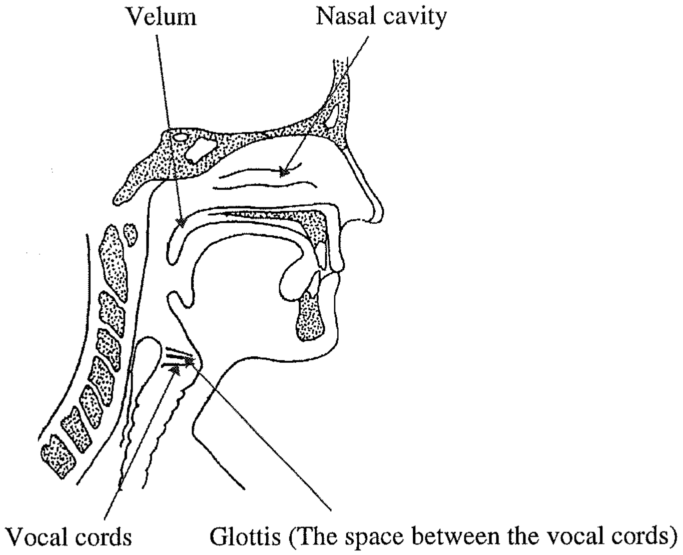


**Figure 1.1. The vocal tract.**

(Diagram from [Holmes, 1988, p.19].)

Approximately four thousand languages are spoken in the world today, each of which encompasses a finite set of elements termed phonemes. These form the basic components of all its words and are defined as the smallest unit in speech where substitution of one with another changes the meaning of the word. For example the words 'car' and 'far' differ in the initial phoneme and the words 'cat' and 'cap' in the final phoneme. In the (British) English language there are approximately forty-four phonemes[2], listed in Figure 1.2 using the International Phonetic Alphabet notation (IPA). The Hawaiian language in comparison has only thirteen.

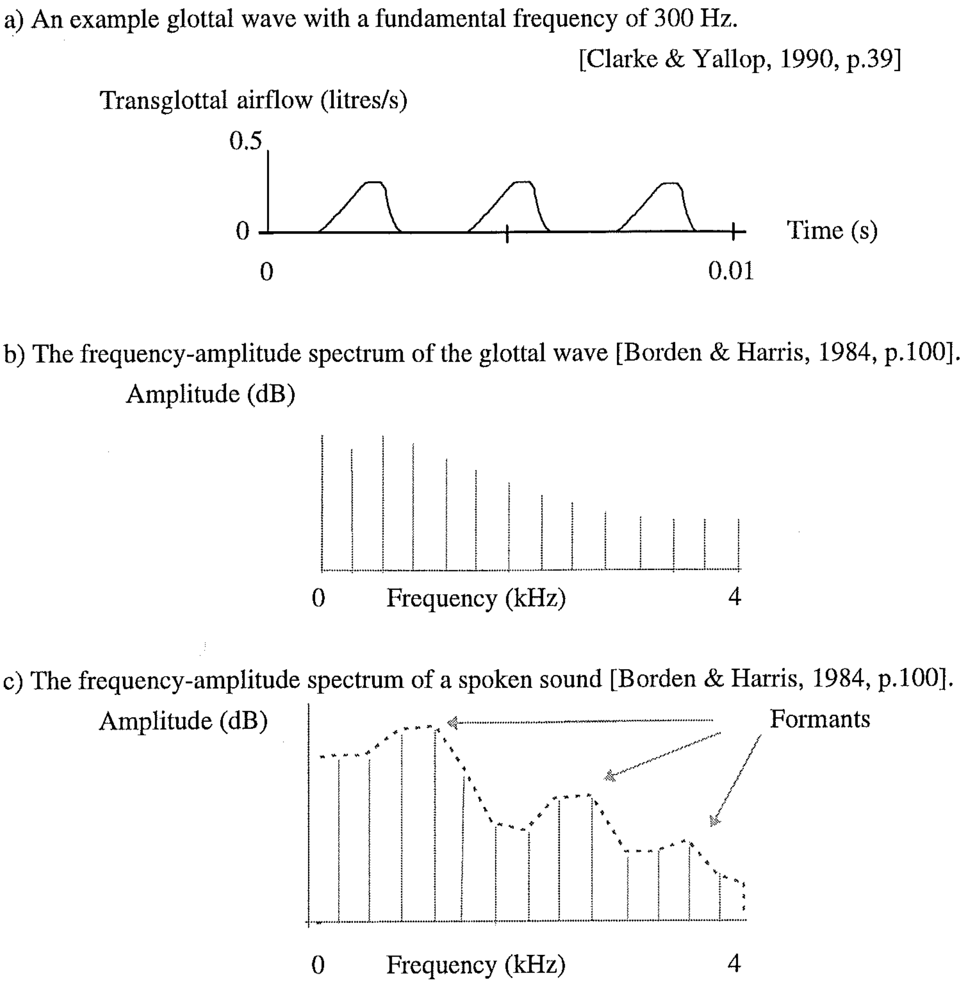
[2] This varies slightly depending on the method of categorisation used.

| Phoneme | Example | Phoneme | Example |
|---|---|---|---|
| i | heed | v | verve |
| ɪ | hid | θ | thick |
| eɪ | hayed | ð | those |
| ɛ | head | s | cease |
| æ | had | z | pizzaz |
| ɑ | father | ʃ | mesh |
| ɔ | hawed | ʒ | measure |
| eʊ | hoed | h | heat |
| ʊ | hood | m | mom |
| u | who'd | n | noon |
| ɜ | heard | ŋ | ringing |
| ə | *a*head | l | lulu |
| ʌ | bud | l | batt*le* |
| ɑɪ | hide | m | botto*m* |
| ɑʊ | how'd | n | butto*n* |
| ɔɪ | boy | ʔ | Glottal stop |
| p | pop | w | wow |
| b | bob | j | yoyo |
| t | tug | r | roar |
| d | dug | tʃ | church |
| k | kick | dʒ | judge |
| g | gig | | |
| f | fit | | |

**Figure 1.2. The English phonemes written in IPA symbols with examples.**

The phonemes within a language may be split into two broad categories: sonorants and non-sonorants. The former include those phonemes featuring a voiced quality such as /ɪ/ and /i/. The latter encompass those phonemes with no tonal or voiced quality such as /k/ and /f/. These are produced by the vocal tract in differing manners.

The excitation of sonorants is caused by passing air from the lungs through the vocal cords, causing them to vibrate. This produces a quasi-periodic glottal wave (Figure 1.3a) whose frequency spectrum features a roll-off of typically –12 dB per octave (Figure 1.3b). The vocal tract acts as a frequency filter which serves to enhance certain frequencies in this spectrum termed formants and to diminish others (Figure 1.3c). The filter characteristics are a function of the vocal tract's shape and by changing its dimensions the locations of the formants are modified. The shape of the vocal tract is determined chiefly by the lips and the tongue. In Figure 1.4 the position of the tongue and its relation to certain phonemes is illustrated.



a) An example glottal wave with a fundamental frequency of 300 Hz.

[Clarke & Yallop, 1990, p.39]

b) The frequency-amplitude spectrum of the glottal wave [Borden & Harris, 1984, p.100].

c) The frequency-amplitude spectrum of a spoken sound [Borden & Harris, 1984, p.100].

**Figure 1.3. Illustrations of the original and filtered glottal wave.**

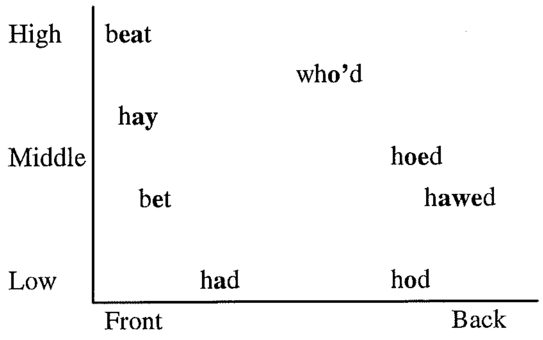


**Figure 1.4. Location of the high point of the tongue for various vowels.**
(Adapted from [Clark & Yallop, 1990, p.68].)

Formants were first observed in 1930 by Sir Richard Paget while listening to a plain chant at Magdalen College, Oxford [Paget, 1930]. Paget observed that some vowels for example /i/ had a 'higher sound' than others for example /ʌ/. Having trained his ears to detect these features, Paget produced a 'vowel resonance chart' summarising for the first time the formant properties of various phonemes (Figure 1.5).

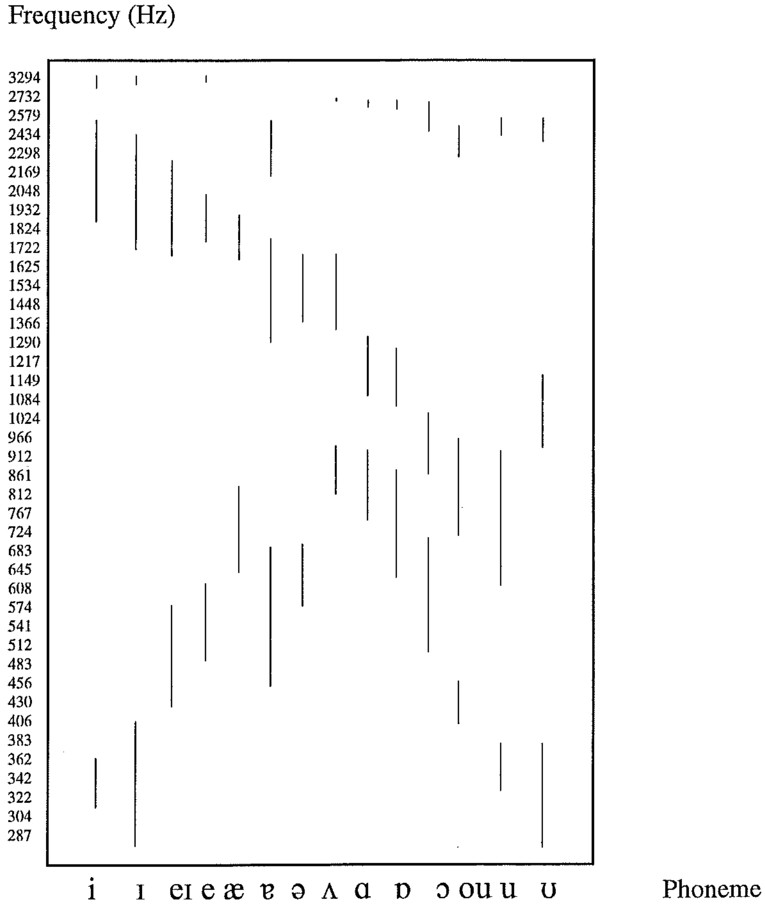


**Figure 1.5. Paget's vowel resonance chart.**
(Diagram from [Linggard, 1985, p.11].)

The second category of phonemes concerns non-sonorants. These encompass phonemes created not from the vibrating vocal cords but from constrictions formed in the vocal tract. The fricatives /s/ and /f/ for example are generated by passing air through constrictions in the vocal tract causing the turbulence which we hear. If the constriction is completely closed allowing pressure to build up before being released, plosives are formed such as /p/ and /k/. In addition some phonemes require both voicing and a constriction in the vocal tract such as /z/ and /v/. These are termed voiced fricatives[3].

When spoken, neighbouring phonemes yield a considerable influence on each other. This property is termed coarticulation and has proved to be a perennial problem in speech synthesis. A second important feature is the manner in which the emphasis, duration and pitch of successive syllables vary. This is known as prosody and must be simulated too if convincing speech synthesis is to be achieved.

## i. Low-level synthesis models

The multitude of low-level synthesis models produced over the last two hundred years may be grouped into three broad categories: articulatory models, spectral modelling and source-filter models. These are reviewed in turn, one of which is selected for inclusion in my singing synthesizer.

### *Articulatory models*

Articulatory synthesizers operate by modelling either literally or through simulation the physical shape of the vocal tract. A simulation of the glottal wave is passed through the model which enhances particular frequencies producing formants. Changing the resonant properties of the model causes the formants to vary, and in this manner produces speech. This approach is an attractive one as it advocates the use of control data which is both tangible and physically meaningful. Many researchers believe this to hold the ultimate solution to successful speech and singing synthesis, the principal argument being that once the evolving shape of the vocal tract is correctly described the perennial problem of coarticulation will be solved. I too believe that this will eventually prove true.

The earliest articulatory synthesizers may be traced back to the eighteenth century. In 1769 the following questions were posed at the Imperial Academy of St. Petersburg:

> i. what are the differentiating properties of the vowels a, e, i, o and u? and
> ii. can these sounds be synthesized?

In response a Russian professor named Kratzenstein constructed five tubes which when excited by a blown reed, emitted sounds bearing perceptual similarities to the five vowels (Figure 1.6).

[3] For a detailed review of phonemic categorization see [Crystal, 1980; Ladefoged, 1993].

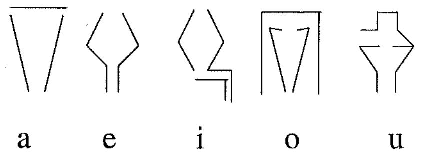


**Figure 1.6. Kratzenstein's resonators.**
(Diagram from [Linggard, 1985, p.9].)

The tubes were rigid and able to produce only static sounds. Similar models were incorporated much later into a musical keyboard (Figure 1.7) forming an early singing synthesizer [Paget, 1930].

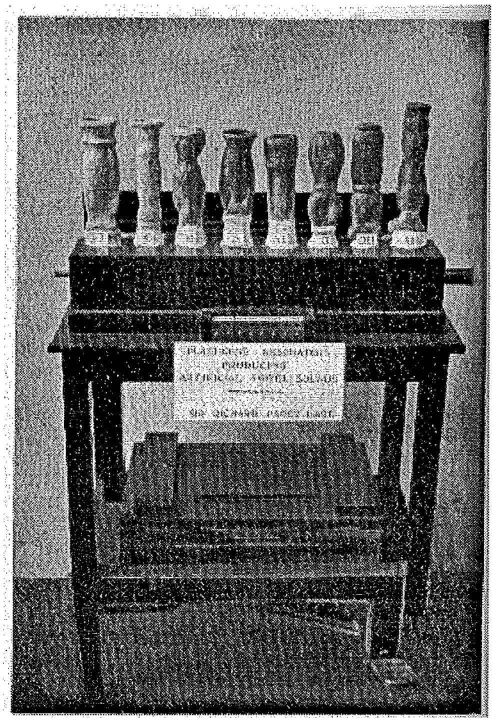


Fig 4. PLASTICINE MODELS ON LORD RAYLEIGH'S ORGAN

**Figure 1.7. Lord Rayleigh's organ.**
(Photo from [Paget, 1930, p.34].)

In 1791 Von Kempelen proposed a model which emulates the dynamic shape of the vocal tract (cited in [Flanagan, 1972, p.206]). Within the model are bellows which when compressed, force air through a reed. This produces a sound wave which passes through a tube whose shape may be manipulated with the hands, causing its resonant frequencies to vary. A second air passage allows the creation of fricatives such as /ʃ/ ('sh') and /s/ ('ss'). It was reported that with skilful control comprehensible speech could be produced in both French and Italian. Von Kempelen's machine was the result of many previous attempts of which he said 'a powerful horse would have had difficulty in pulling'. He claimed to have designed this model thirty years previously but had been ignored because of an earlier faux pas: a chess playing machine he had designed and flaunted some years before was found to consist in reality of a small cupboard in which was concealed a legless chess player. If as claimed, his model of the vocal tract was truly built in 1761, it almost certainly marks the birth of speech synthesis. Figure 1.8 illustrates a reconstruction of Von Kempelen's system by Wheatstone in 1879. This was to impress many speech researchers of the time including a young Alexander Bell.

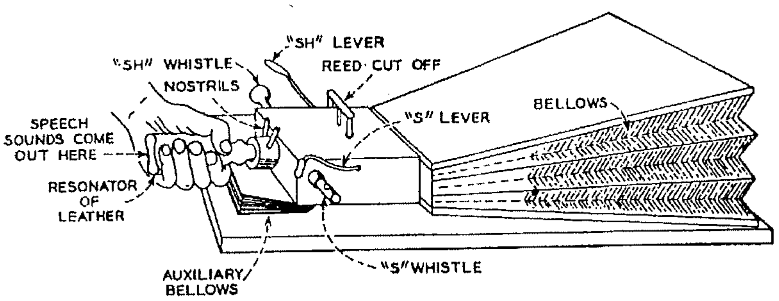


**Figure 1.8. Wheatstone's construction of Von Kempelen's speaking machine.**
(Diagram from [Flanagan, 1972, p.206].)

A similar design was seen in Faber's pipe organ, which reportedly was able to sing 'God save the Queen' (Figure 1.9).

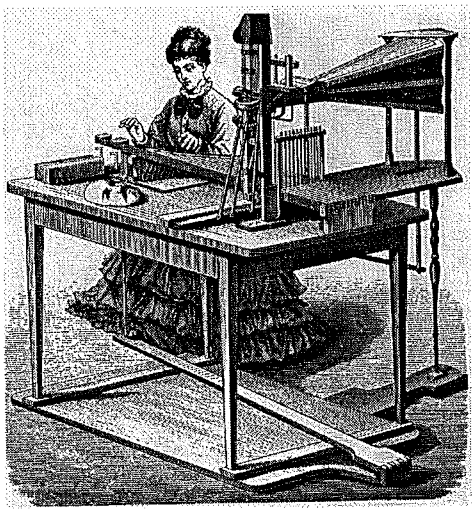


**Figure 1.9. Joseph Faber's speech organ.**
(Taken from the Museum of Speech and Synthesis Web page
http://mambo.ucsc.edu/psl/smus/smus.html.)

In 1937 a model of the vocal tract named the Riesz synthesizer was constructed which could be manipulated with one of eight control keys along its upper surface (cited in [Flanagan, 1972, p.208]) (Figure 1.10). Compressed air was passed through a reed and into the model. A skilled operator could then depress the keys in a controlled manner, modifying the resulting wave to produce comprehensible speech.

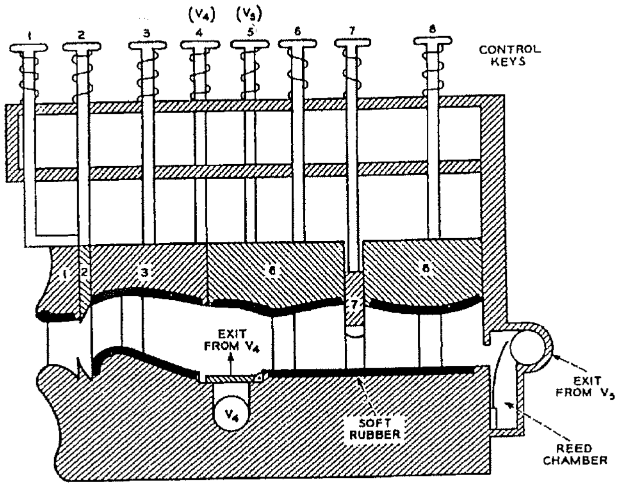


**Figure 1.10. The Riesz synthesizer.**
(Diagram from [Flanagan, 1972, p.208].)

From 1940 X-ray techniques allowed phoneticians to view the mechanisms of the vocal tract in previously unattainable detail. From these pictures the area function of the vocal tract could be estimated and used as descriptive data for articulatory synthesizers. This technique, however, was and still is prone to problems. Any participant can be subjected only to minimal amounts of X-ray dosage, and the resulting pictures are rather blurred (Figure 1.11).

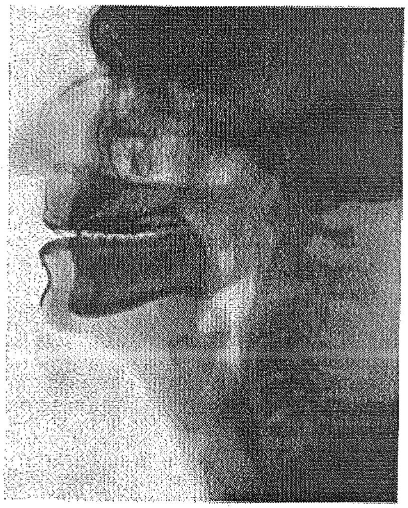
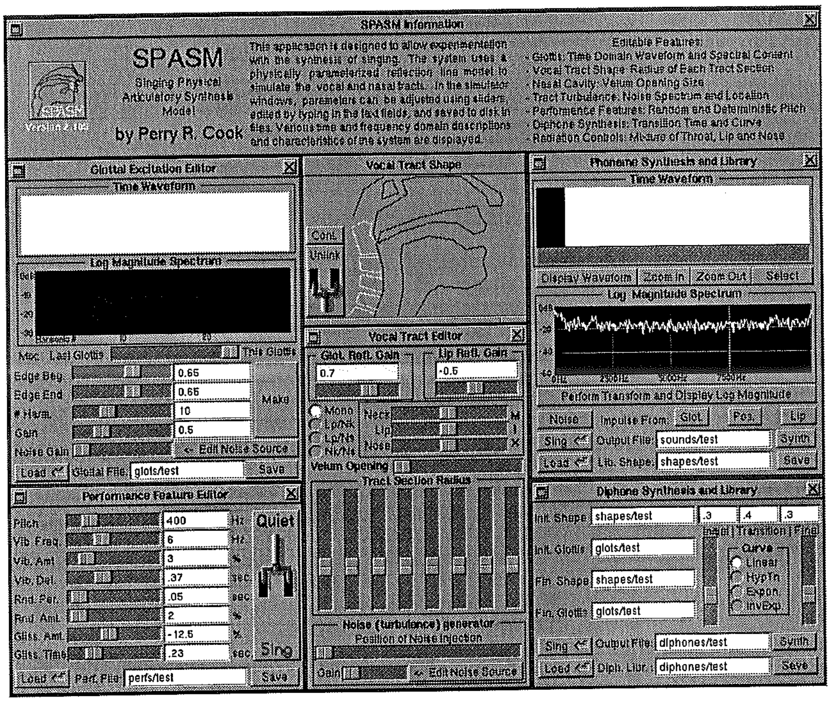


**Figure 1.11. An X-ray of the vocal tract.**
(Photo from [Flanagan, 1972, p.188].)

Direct physical modelling of the vocal tract lost impetus following the Riesz synthesizer for two reasons: firstly, constructing a precise and flexible model of the vocal tract was and remains highly problematic; secondly, the advent of electrical circuitry and transmission line theory led to its simulation in the electrical domain [Dunn, 1950; Stevens *et al.*, 1955; Flanagan, 1972, p.272–276; Linggard, 1985, p.38–66]. These were followed in 1958 by an articulatory synthesizer which allowed continuous speech to be produced through dynamic control of adjoining electrical networks [Rosen, 1958].

From approximately 1985 the advent of fast digital computers led to the simulation of articulatory synthesis in the digital domain [Liljencrants, 1985]. These have been followed by a number of singing synthesizers including Pavarobotti[4] [Titze & Story, 1993] (audio examples one and two) and the SPASM synthesizer[5] [Cook, 1990; Cook, 1993a; Cook, 1993b; Cook, 1993c] (audio examples three and four). The latter is implemented on a Next machine renowned for excellent user interface designs. Using a graphical interface, the user is able to modify the glottal wave, vocal tract dimensions and fundamental frequency to produce various phonemes and diphones under prosodic control (Figure 1.12).

[4] Details may be found on the Web page http://www.shc.uiowa.edu/fun/pavarobotti/pavarobotti.html.
[5] Details may be found on the Web page http://www.cs.princeton.edu/~prc/SingingSynth.html.



SPASM Information

SPASM

Singing Physical Articulatory Synthesis Model

by Perry R. Cook

Version 2.105

This application is designed to allow experimentation with the synthesis of singing. The system uses a physically parameterized reflection line model to simulate the vocal and nasal tracts. In the simulator windows, parameters can be adjusted using sliders, edited by typing in the text fields, and saved to disk in files. Various time and frequency domain descriptions and characteristics of the system are displayed.

Editable Features:

- Glottis: Time Domain Waveform and Spectral Content
- Vocal Tract Shape: Radius of Each Tract Section
- Nasal Cavity: Velum Opening Size
- Tract Turbulence: Noise Spectrum and Location
- Performance Features: Random and Deterministic Pitch
- Diphone Synthesis: Transition Time and Curve
- Radiation Controls: Mixture of Throat, Lip and Nose

Glottal Excitation Editor

Edge Beg. 0.65
Edge End 0.65
# Harm. 10
Gain 0.5
Glottal File: glots/test

Vocal Tract Editor

Glot. Refl. Gain 0.7
Lip Refl. Gain -0.6

Performance Feature Editor

Pitch 400 Hz
Vib. Freq. 6 Hz
Vib. Amt. 3 %
Vib. Del. .37 sec.
Rnd. Per. .05 sec.
Rnd. Amt. 2 %
Gliss. Amt. -12.5 %
Gliss. Time .23 sec.
Perf. File: perfs/test

Phoneme Synthesis and Library

Output File: sounds/test
Lib. Shape: shapes/test

Diphone Synthesis and Library

Init. Shape shapes/test .3 .4 .3
Init. Glottis glots/test
Fin. Shape shapes/test
Fin. Glottis glots/test
Output File: diphones/test
Diph. Libr.: diphones/test

**Figure 1.12. A screen dump of the SPASM interface.**

An important factor in all articulatory synthesizers is the quality of the sound source itself. To attain realistic speech the sound source should be as similar to a natural glottal wave as possible. The task of deriving a description of the glottal wave has, however, proved highly problematic [Klatt, 1987a, p.745–746]. The early synthesizers used coarse approximations such as saw-tooth waveforms and filtered impulse trains whose spectral properties bore reasonable similarities to those of glottal waves (Figure 1.13).

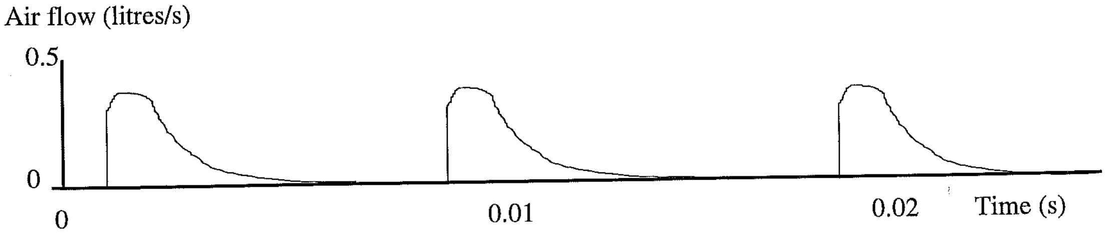


**Figure 1.13. A filtered impulse train used to emulate the glottal wave.**
(Diagram from [Klatt, 1987a, p.744].)

Further research may be seen in [Flanagan & Langraf, 1968; Titze, 1974; Titze, 1984; Holmes, 1973; Rosenberg, 1971; Rothenberg *et al.*, 1975; Fant *et al.*, 1985; Sundberg *et al.*, 1990] in which are proposed techniques including the use of three-dimensional graphical models and the inverse-filtering of vowels.

In summary, articulatory synthesis is an attractive option as the control parameters describe tangible properties. There are however a number of problems:

- obtaining accurate measurements of the dynamic vocal tract is problematic;
- the singer must be available for analysis (and therefore still alive);
- plane-wave propagation is assumed, limiting the synthesis sampling rate to under approximately 6 kHz [Holmes, 1988, p.19];
- finding an accurate description of the glottal wave is problematic;
- the shape of the vocal tract is highly complex and therefore difficult to simulate.

With the advent of magnetic resonance imaging (MRI), the first problem is somewhat ameliorated, allowing precise imaging of the vocal tract to take place [Narayanan *et al.,* 1995]. As this technology continues to improve, detailed mapping of the vocal tract's movements will be possible. The remaining problems, however, persist. High-quality output from articulatory synthesis for extensive amounts of textual input, seems unattainable at present.

## *Spectral modelling*

Spectral modelling synthesizers generate speech waves by emulating their required frequency spectra without attempting to model the system by which they are produced (in this case the vocal tract). The tool used most commonly to derive such frequency spectra is the spectrograph. This is perhaps the most commonly used application in sound analysis laboratories and warrants a brief description before a review of spectral modelling is given.

### ***The spectrograph***

The spectrograph was designed in 1946 with the purpose of encoding the frequency properties of an audio wave in a printed format [Koenig *et al.,* 1946]. The design of an early spectrograph termed the Model D Sonagraph is illustrated in Figure 1.14.

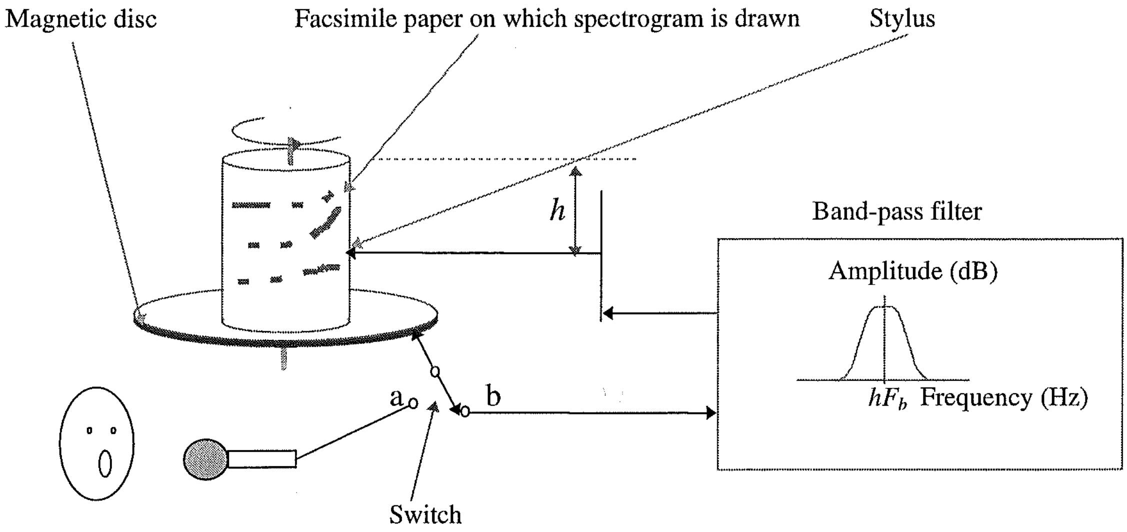


**Figure 1.14. The Model D Sonagraph.**
(Adapted from [Flanagan, 1972, p.150].)

The output of the spectrograph is a spectrogram: a graph of time against frequency in which is illustrated the dynamic frequency properties of a wave. This is produced in two steps. With the switch in position **a**, a wave *W* is recorded onto the rotating magnetic disc. The switch is then placed in position **b**. Fixed to the magnetic disc is a cylinder on which is attached a sheet of heat-sensitive paper. When revolved, wave *W* is played through a band-pass filter whose output controls the heat of a stylus touching the paper at a height $h$. The central frequency of the filter is equal to $hF_{step}$ Hz where $F_{step}$ is set typically to 300 Hz. Thus in one revolution the filter will cause the stylus to mark on the paper locations where frequency components around $hF_{step}$ Hz are present. The value of $h$ is increased for each of many revolutions. On completion the heat-sensitive paper is marked with the spectrogram of wave *W*.

If the bandwidth of the filter is narrow (45 Hz), its time response will be large. Thus the filter will respond only to a narrow range of frequencies but with temporal inertia. If the bandwidth of the filter is wide (300 Hz), its time response will be small. The filter will then have a rapid response but to a wide range of frequencies. These produce spectrograms, termed narrow-band and wide-band respectively, in which there is a trade-off between frequency and temporal resolution. Digital simulations of the spectrograph are found in virtually all current sound analysis software. Figure 1.15 illustrates wide- and narrow-band spectrograms of the word ‘ago’ in which their contrasting styles may be observed. The wide-band spectrogram has a good temporal resolution but poor frequency resolution. This allows the instance at which the phoneme /g/ is pronounced to be located and reveals the formant properties. The narrow-band spectrogram displays better frequency resolution allowing individual harmonics to be observed but with a reduced temporal resolution.

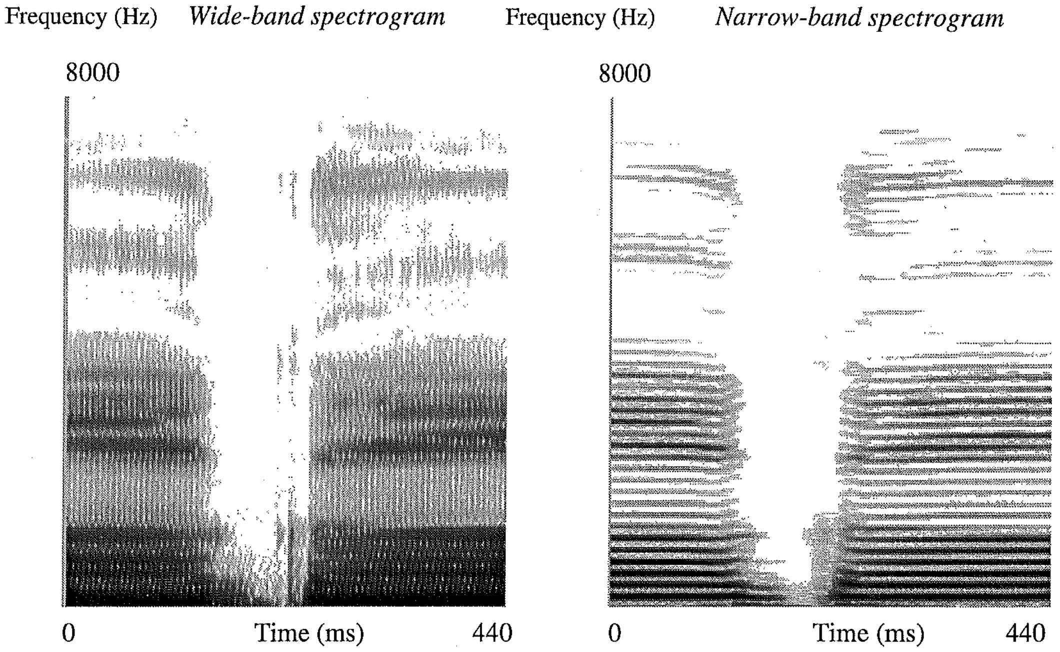


**Figure 1.15. Wide- and narrow-band spectrograms of the word 'ago'.**

The first spectral modelling synthesizers were developed over one hundred years before the spectrograph was invented, following work by Fourier. In his research, Fourier observed that periodic waves could be described as a sum of sinusoids. The (quasi-) periodic waves observed in sonorants could therefore be synthesized by summing sinusoids generated from a number of sound sources. This led to several ingenious designs incorporating tuning forks, sirens and pipes (see [Flanagan, 1972, p.207]).

Fourier's work remains influential in current research, in which digital computers are used to analyse and synthesize audio waves. The two relations, the discrete Fourier transform (DFT) and the inverse discrete Fourier transform, allow a wave to be viewed in both its temporal and frequency domains. In digital computers, audio waves are stored as a series of samples and are termed sampled waves. The $n$th sample of the sampled wave $f$, denoted by $f(n)$, describes the magnitude of $f$ at time $n/f_s$, where $f_s$ is the sampling frequency. If $f$ is a sampled wave of length $N$ samples, its discrete Fourier transform $F$ is defined as

$$F(k) = \frac{1}{N}\sum_{n=0}^{N-1} f(n)e^{-2\pi ikn/N} \ ,$$

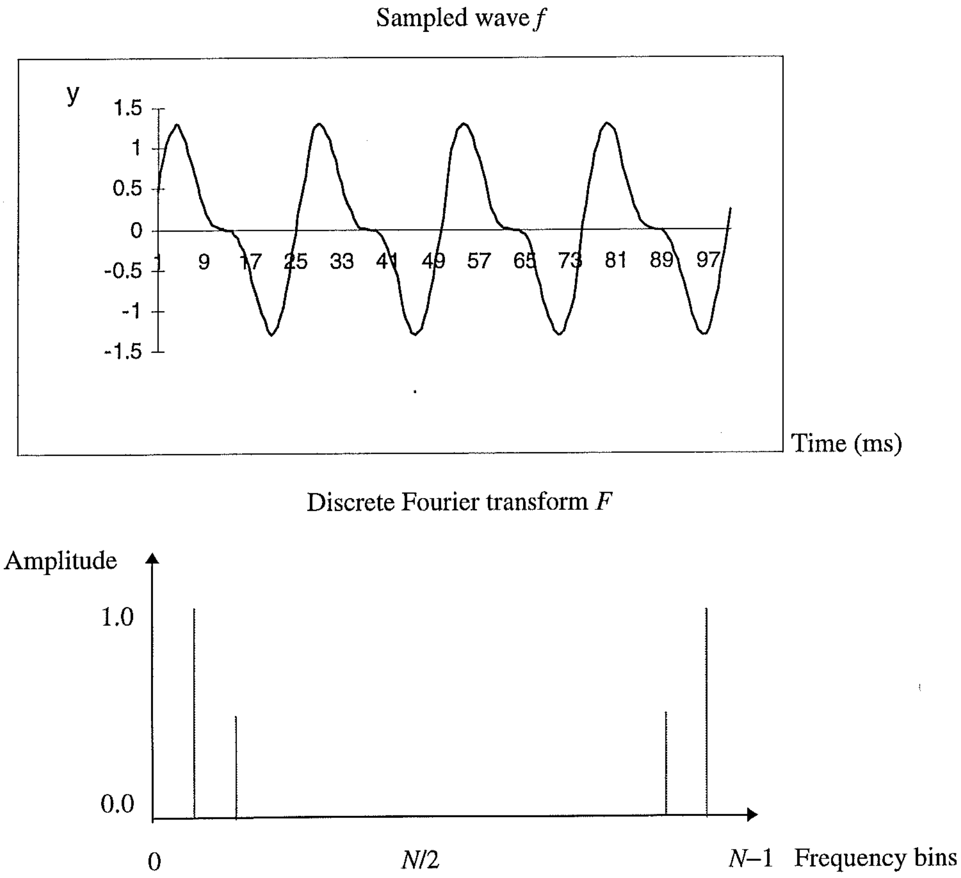
where $0 \leq k < N$.

Conversely, given a discrete Fourier transform $F$ of length $N$ samples, the wave $f$ which it represents is given by

$$f(n) = \sum_{k=0}^{N-1} F(k)e^{2\pi ikn/N} \ ,$$

where $0 \leq n < N$.

An example of a sampled wave $f$ and its discrete Fourier transform $F$ is given in Figure 1.16.


Sampled wave *f*

Discrete Fourier transform *F*

**Figure 1.16. A periodic wave *f* and accompanying Fourier transform *F*.**

### *Spectral modelling using additive synthesis*

The inverse discrete Fourier transform provides a means for deriving the temporal representation of a wave $f$ from its frequency representation (DFT) $F$. A second approach, termed additive synthesis, provides an alternative means for deriving a temporal wave from its frequency representation. If the frequency, amplitude and phase of each component sinusoid of a wave $f$ is known, $f$ may be resynthesized directly using the additive synthesis equation

$$f(n) = \sum_{i=1}^{P} a_i \sin(nf_i + \varphi_i)\,,\ 0 \le n < N\,,$$

where $a_i$, $f_i$ and $\varphi_i$ are the amplitude, frequency and phase of the $i$th sinusoid, respectively,

$N$ is the length of the wave and

$P$ is the number of component sinusoids.

Unlike the inverse Fourier transform, the additive synthesis equation allows the sinusoids' amplitudes, frequencies and phases to be described as a function of time and thus change at any point along the wave. The previous equation then becomes

$$f(n) = \sum_{i=1}^{P} a_i(t) \sin(nf_i(t) + \varphi_i(t))\,,\ 0 \le n < N\,,$$

where $a_i(t), f_i(t)$ and $\varphi_i(t)$ are functions of time and

$t = n / f_s$.

This provides an extremely powerful technique for audio synthesis, particularly for the synthesis of voiced sounds. The computational complexity of the additive synthesis algorithm is directly related to the number of sinusoids being synthesized. Additive synthesis is therefore suitable for voiced sounds, in which there are a low number of sinusoids (e.g. less than 100) but not suitable for unvoiced sounds, for which another approach, for example inverse discrete Fourier transforms, should be used (see Chapter 4).

The means by which the functions $a_i(t), f_i(t)$ and $\varphi_i(t)$ are derived, are explained in detail in Chapter 4 and in [Serra & Smith, 1990; Fitz & Haken, 1997] (see also the Web page http://www.iua.upf.es/~sms/). A novel application of additive synthesis is reviewed in [Depalle *et al.*, 1994] and concerns the synthesis of a castrato's voice for the internationally acclaimed film 'Farinelli'[6]. In addition a speech synthesizer has been developed recently at KTH [Granqvist, 1996] based upon this technique.

### *The Pattern Playback synthesizer*

To conclude the review of spectral modelling, it is appropriate to mention the Pattern Playback synthesizer [Cooper, 1950]. Additive synthesis in the digital domain is a computationally intensive operation. Only recently has it been possible to perform this in real time. In 1950 the Pattern Playback synthesizer was designed which could synthesize the data represented in a spectrogram immediately. This system contains a transparent roll on which a spectrogram is painted (Figure 1.17). A lamp and optical system produce together a narrow fan-shaped beam of light which is passed through a 'tone wheel'. This consists of a rotating disc on which are printed concentric annular patterns. These represent a fundamental frequency and successive harmonics. As the light is passed through the disc it is modulated along its width by these patterns. The modulated beam then passes through the painted spectrogram onto a photocell which amplifies the received signal and converts it into sound. This system allows users to paint spectrograms of sounds which may then be synthesized and heard immediately (see also [Schott, 1948]).

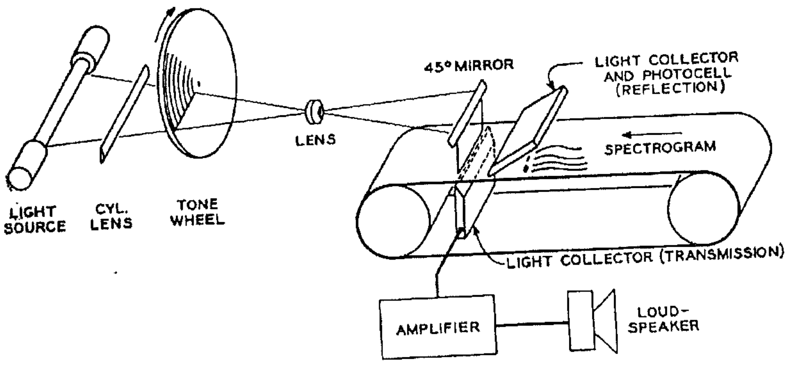


**Figure 1.17. The Pattern Playback synthesizer.**
(Diagram from [Flanagan, 1972, p.213].)

[6] Details of this system may be found on the Web page:
http://www.ircam.fr/produits-real/multimedia/farinelli-e.html.

## *The source-filter theory of speech production and ensuing models*

In 1960 an influential thesis entitled 'Acoustic theory of speech production' was published [Fant, 1960] (see also [Dunn, 1950]). This formalised the notion that the production of speech may be viewed as the excitation of a filter by one or more sound sources. Two sound sources are considered: a periodic wave for emulating the glottal sound source and a noise wave for emulating turbulence. The filter modifies these sounds in a manner which reflects the filtering properties of the vocal tract (Figure 1.18). The source-filter principle forms the basis of many speech synthesis models including vocoders, formant synthesizers and LPC synthesizers. These are reviewed in turn.

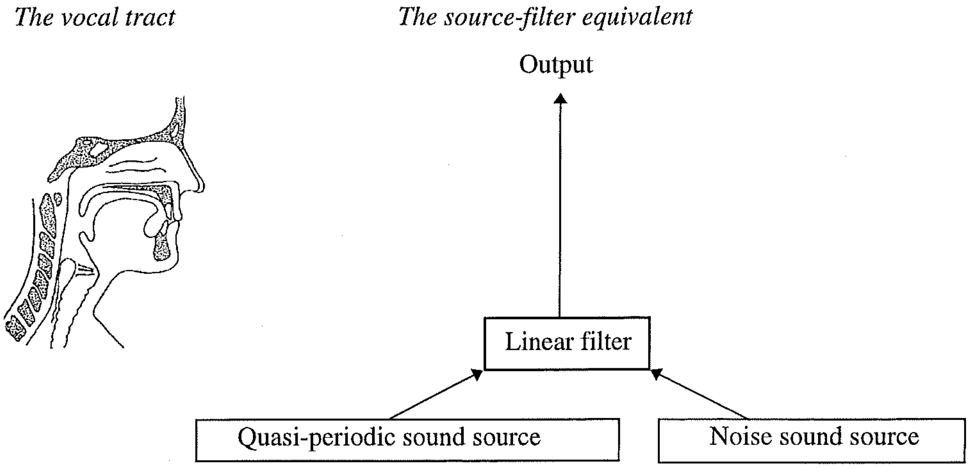


**Figure 1.18. The source-filter model.**

### *The vocoder*

The vocoder or 'VOice enCODER' is regarded by many as the first major step in continuous speech synthesis [Dudley, 1939]. Its main purpose is to transmit speech at a low bandwidth, a feat which is achieved by exploiting the source-filter model of speech production. The vocoder consists of an analysis and a resynthesis section (Figure 1.19).

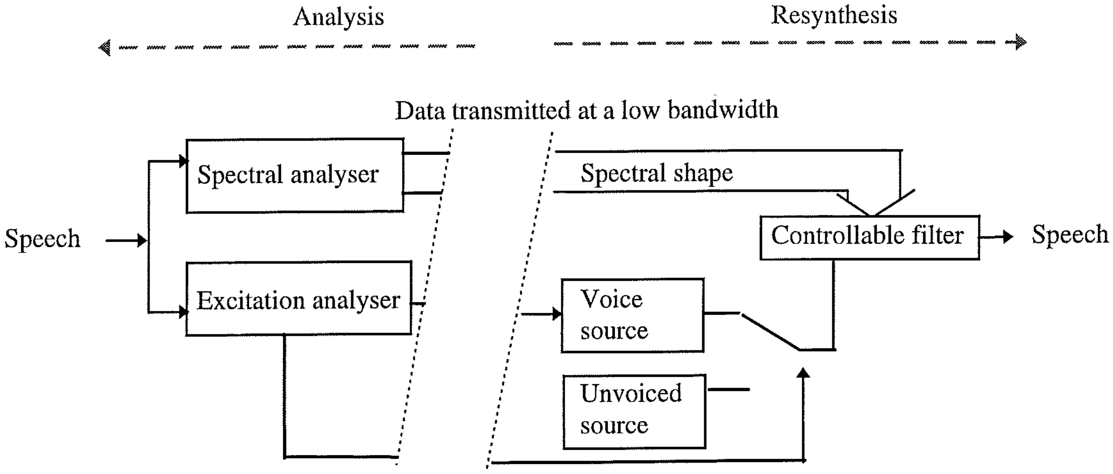


**Figure 1.19. The general vocoder structure.**
(Diagram from [Holmes, 1988, p.60].)

The analysis part encodes the speech wave as two components: a description of the excitation sound source and a description of the overall frequency spectra. The original vocoder, termed the channel vocoder, encodes the excitation as either a voiced flag with fundamental frequency *P* or as an unvoiced flag. The frequency spectrum of the wave is encoded by observing the response of a ten-channel filter bank. This data is transmitted to another vocoder which then resynthesizes an approximation of the original wave. The excitation data determines which of two sound sources is to be used: a noise source or a periodic source with a fundamental frequency *P*. This is passed through a ten-channel filter bank controlled by the output of the original filter bank.

The vocoder formed the basis of the 'Voder' synthesizer which proved to be the main attraction at the New York World Trade Fair in 1939 (Figure 1.20). This features manual control of the sound source and filter parameters, allowing skilled operators to synthesize novel utterances in real time.

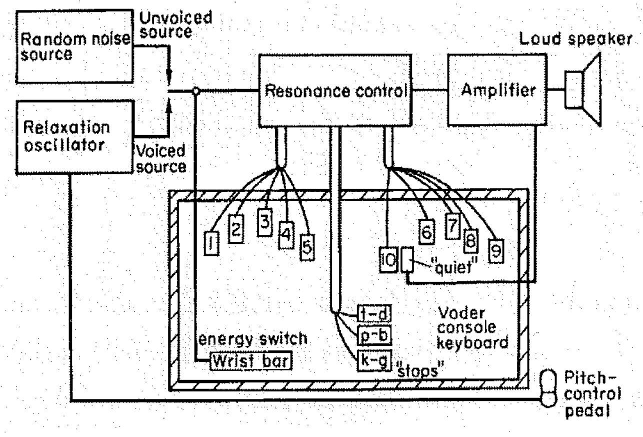


**Figure 1.20. The Voder.**

(Diagram from [Flanagan, 1972, p.212].)

### *Formant vocoders*

Formant vocoders operate on a similar principle to the channel vocoder but allow a lower transmission bandwidth. As before, the speech is decomposed into descriptions of the excitation source and spectral shape; however, the vocoders differ in their spectral encoding strategy. Whereas the channel vocoder encodes the frequency spectra as the output of a filter bank, the formant vocoder derives the locations and bandwidths of formants within the spectra.

This is advantageous as the formant positions are a function of the vocal tract's shape which changes only at a relatively slow rate: the tongue and jaw are bodies with mass and momentum moved by muscles with a finite force. The formants observed in speech therefore move relatively slowly. It is found that acceptable speech resynthesis may be attained by transmitting formant positions at typically one hundred times a second. This data may be received by a formant vocoder and used to control electronic resonators which act as filters on an excitation source.

The first speech to be synthesized using formant synthesis was in 1922 (see [Flanagan, 1972, p.211]). This system comprised of a buzzer which excited two static resonators. This was followed fourteen years later with a similar system comprising of four static resonators and a buzzing sound source. Only in 1950 was formant synthesis used in conjunction with the vocoder's system of analysis and accompanying resynthesis to produce continuous speech [Munson & Montgomery, 1950]. This system used crude formant and pitch estimators and was a catalyst to the development of improved formant and pitch tracking techniques [Holmes, 1958; Dolanksy, 1955].

Formant synthesis has remained an active area of research leading to, amongst others, the OVE (Orator Verbis Electris) formant synthesizer [Fant, 1953]. This consists of cascade filters, the two lowest of which are controlled by a mechanical arm (Figure 1.21). Shortly afterwards an electronic conversation was performed between a similar synthesizer PAT (Parametric Artificial Talker) and OVE at the research centre the Massachusetts Institute of Technology. This was perhaps the first ever synthesized discussion [Lawrence, 1953].

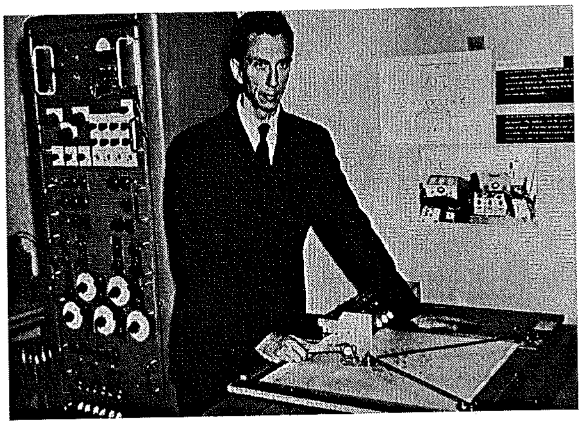


**Figure 1.21. Gunnar Fant and the formant synthesizer.**

(Taken from the Museum of Speech and Synthesis Web page http://mambo.ucsc.edu/psl/smus/smus.html.)

The advent of digital computers in the early 1960's allowed analogue formant resonators to be replaced by digital filters. The earliest digital formant vocoder was probably the OVE II, the successor to the OVE, developed in 1962 [Fant & Martony, 1962]. This was followed by many others [Coker & Cummiskey, 1965; Gold & Rabiner, 1968; Liljencrants, 1967].

Perhaps the most important step in formant synthesis was taken in 1972 when Holmes synthesized the phrase "I enjoy the simple life" so similar to the original as to be perceptually identical. This was the first reported time a synthesized phrase had sounded as realistic as the original. His model was to be produced later as a microchip [Quarmby & Holmes, 1984]. Other products arising from formant based synthesis include DECtalk [Klatt, 1987b] and Infovox[7].

Two contrasting models of formant synthesizers have emerged in which the filters are placed either in parallel or in series. The former allows more flexibility but requires the additional specification of the formants' amplitudes. The contrasting strengths of cascade and parallel systems is described in [Holmes, 1983] and a combination of both systems can been found in several synthesizers including the Klatt synthesizer [Klatt, 1980].

Three formant synthesizers have been employed in singing synthesis: the MUSSE synthesizer [Larsson, 1977] (audio examples five and six), the CHANT synthesizer[8] [Rodet *et al.*, 1984; Rodet, 1985] and Pabon's synthesizer [Pabon, 1993] (audio example seven). The MUSSE (MUsic and Singing Synthesis Equipment) synthesizer is based upon an analogue formant model containing five filters. The pitch is input from a musical keyboard to which a variable vibrato is applied. The formant locations and vibrato specifications are input either manually or from a computer. In 1988 this was superseded by a digital version named MUSSE–DIG, modelled with seven filters and a second branch for producing consonants (Figure 1.22) [Berndtsson & Sundberg, 1993; Carlsson & Neovius, 1990; Carlsson *et al.*, 1991]. This combined features from the earlier MUSSE synthesizer and the OVE family of synthesizers.

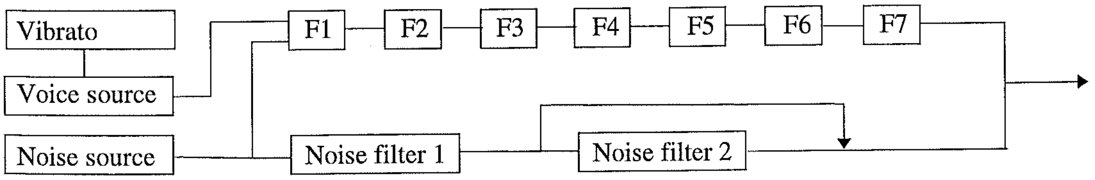


**Figure 1.22. The MUSSE–DIG synthesizer.**
(Diagram from [Berndtsson, 1995, p.3].)

With present day computer power this is capable of operating in real time. Ensuing additions to the MUSSE–DIG synthesizer have included an option to apply small

[7] See the Web page http://www.communique.se/polarprint/infovox.html.

[8] See the Web page http://www.ircam.fr/produits-real/logiciels/chant-e.html.

scale deviations to the pitch path termed 'jitters' [Ternström & Friberg, 1989] and further improvements to the glottal wave [Pabon, 1993].

Formant research specific to the singing voice has also proved a popular area of research, in which properties such as the singer's formant and formant tuning in singing have been examined [Berndtsson & Sundberg, 1994; Carlsson & Sundberg, 1992; Pillot, 1995; Sundberg, 1968; Sundberg, 1981; Sundberg, 1995; Berg *et al.*, 1995; Johnston & Scherer, 1995].

Extensive research into formant synthesis has led to a wealth of data in this field. In addition it is possible to buy formant synthesizers 'off the shelf'. However there are a number of disadvantages to this approach:

- obtaining accurate formant tracks automatically is troublesome [Parsons, 1986, p.210]. Problems arise for example when resolving two formants whose central frequencies lie close to each other. It is difficult too to determine the formant positions within high-pitched notes whose harmonics lie far apart and whose frequency spectra are therefore sparsely defined;
- eliciting an accurate description of the glottal wave is troublesome [Klatt, 1987a, p.745–746];
- most formant synthesizers assume a non-branching vocal-tract model, making them suitable for voiced non-nasal sounds such as /ɪ/ but inappropriate for nasal sounds such as /m/ in which there are also anti-resonances. More complex formant synthesizers [Klatt,1980] do model these anti-resonances, leading to improved results.

### *Linear predictive coding (LPC) vocoders*

Linear predictive coding [Schroeder, 1985], abbreviated often to LPC, was first seen in the domain of speech synthesis in 1968 [Itakura & Saito, 1968; Atal & Hanauer, 1971]. This proved particularly popular in the vocoder system as, although computationally intensive, the analysis process is entirely automated. With the enhanced performance of later computers, LPC has become an extremely popular technique in speech analysis, synthesis and resynthesis.

Linear predictive coding utilises the property that the *n*th sample in a wave *s* produced by a resonating system may be predicted by forming a linear weighting of its past samples. That is, the predicted value of $s(n)$ labelled $s'(n)$ is given by

$$s'(n) = a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p),$$

where $a_1$ .. $a_p$ are termed the linear prediction coefficients. The prediction error is then

$$e(n) = s(n) - s'(n).$$

It is possible to determine separate prediction coefficients for each sample *n* such that the error term $e(n)$ is precisely zero. However, this provides no advantages as more data would be needed to encode the coefficients than is needed to encode $s(n)$ directly. The advantage of LPC is that the *p* predictor coefficients may be determined so that $e(n)$ is optimally small over a series of *M* consecutive samples, where *M* is much larger than *p*. This results in a large saving in data bandwidth. *M* samples of the wave may then be approximated by taking the prediction error signal and the predictor coefficients using the equation

$$s(n) = e(n) + \sum_{i=1}^{p} a_i s(n-i).$$

For voiced sounds the error signal $e(n)$ approximates a pulse train and for unvoiced sounds it approximates a noise signal (see [Linggard, 1985, p.101]). The LPC vocoder exploits this property. If the currently encoded speech is voiced, the data transmitted by the vocoder are the predictor coefficients and the frequency and amplitude of the idealised pulse train which approximate the error signal. If the sound is unvoiced, an unvoiced flag is set and transmitted. A second vocoder may then resynthesize the wave from the error description and predictor coefficients using the last equation. An example is illustrated in Figure 1.23 based upon one of the earliest LPC vocoders to be produced in 1971 [Atal & Hanauer, 1971].

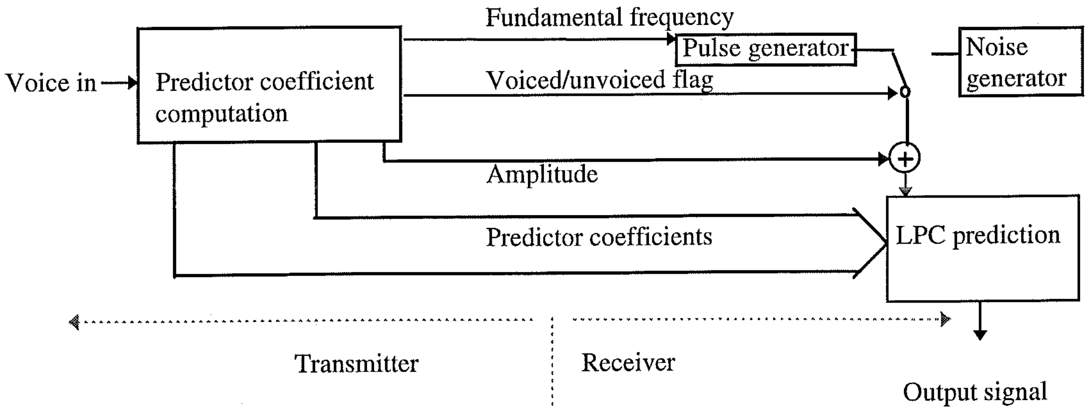


**Figure 1.23. An LPC Vocoder.**
(Adapted from [Holmes, 1988, p.62].)

LPC has proved a popular area in speech synthesis and as a consequence there is a vast amount of research in this field. It is an extremely efficient and highly automated coding technique and is possible to perform in real time. However, the limited quality with which the glottal source is modelled leads to results which are often termed 'buzzy' [Holmes, 1988, p.63].

### *Choice of a low-level synthesis model*

The factors which require consideration when selecting a synthesis model on which to base my singing synthesizer include the following:

the analysis technique should be highly automated,
the presence of the singer should not be required,
real-time processing would be preferable,
a high quality of synthesis is desirable,
control of the vocal quality is desirable and
prosodic control is essential.