Styger, T., & Keller, E. (1994). Formant synthesis.In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges* (pp. 109-128). Chichester: John Wiley.

Formant Synthesis 6

Thomas Styger and *Eric Keller*

Laboratoire d'analyse informatique de la parole (LAIP)
Université de Lausanne, CH-1015 LAUSANNE, Switzerland

A simple approach to the basic "source-filter" model of speech production, as used for synthesis purposes, is presented in this chapter. In the introduction, a short review of the principal synthesis methods is given. Direct synthesis, synthesis using a production model and articulatory synthesis are introduced succinctly. The source-filter model is then described. Topics such as voicing and friction sources, vocal tract filter and lip radiation are summarised. Finally Klatt's formant synthesiser is presented as an application of this theory.

The purpose of the first section of this chapter is to provide an insight into the main speech synthesis techniques that are in common use. The principles of these approaches are exposed to give a short overview and to prepare the reader for the main topic of this chapter, which is formant synthesis. This synthesiser type is based on the source-filter theory of speech production, which will be introduced in the second section. The last section is dedicated to a specific implementation of the above-described theory: The Klatt formant synthesiser.

An Overview of Synthesis Techniques

Automatic speech signal generation on computers is commonly called "speech synthesis". The various generation techniques can be divided into three classes (Calliope, 1989):

— direct synthesis,
— synthesis using a production model,
— vocal tract simulation.

Direct Synthesis

In direct synthesis, the speech signal is generated by direct manipulation of its wave form representation. *Wave form concatenation*;is representative of this synthesis category (Cooper *et al.*, 1951). In this approach, several fundamental periods of pre-recorded phonemes are simply concatenated, or "glued together". The phonemes are then connected to form words and sentences.

This simple technique requires very little computing power or disk space. However, it results in relatively poor quality, which is only acceptable in a few applications, such as in toys. In this method, coarticulation phenomena[1] are not generally taken into account, which is likely to be the reason that speech produced by this technique is of limited intelligibility (Klatt, 1987).

Another direct synthesis technique is the *channel vocoder*;(Dudley, 1939). This procedure generates signals by activating a number of frequency channels, in proportion to the distribution of formant energy in the spectrum (i.e. according to the *vocal tract transfer function*). To simulate the voice fundamental frequency (F_0) source, an evenly-spaced train of impulses is applied. In fricative segments, a noise generator is substituted for this source.

The advantage of this technique is that synthesis proceeds by direct inversion of spectral analysis results, that it is entirely automatic, and that information relating to the excitation is completely independent of information about the transfer function. However, the quality obtained in this manner remains limited, due to the small number of channels used, i.e. because of the relatively gross quantification of the spectrum. For these reasons, the technique is no longer used in current speech synthesis.

Synthesis Using a Speech Production Model

Methods that simulate the speech production mechanisms are mainly based on the source-filter theory. In this approach, a linear filter simulates the vocal tract, which in turn is driven by an adequate source.

Two main techniques belong to this category: The formant synthesiser (Klatt, 1980) and diphone concatenation (O'Shaughnessy *et al.*, 1988), using a linear prediction coding technique (LPC) (Atal and Hanauer, 1971).

---

1 That is, the fact that the same phoneme is produced quite differently, and has rather different spectral characteristics, when appearing in one phonetic context or in another.

Formant Synthesiser

In formant synthesis, the basic assumption is that the vocal tract transfer function can be satisfactorily modelled by simulating formant frequencies and formant amplitudes. The synthesis thus consists of the artificial reconstruction of the formant characteristics to be produced. This is done by exciting a set of resonators by a voicing source or noise generator to achieve the desired speech spectrum, and by controlling the excitation source to simulate either voicing and voicelessness. The addition of a set of anti-resonators furthermore allows the simulation of nasal tract effects, fricatives and plosives.

The specification of about 20 or more such parameters can lead to a satisfactory restitution of the speech signal. The advantage of this technique is that its parameters are highly correlated with the production and propagation of sound in the oral tract. The main current drawback of this approach is that automatic techniques of specifying formant parameters are still largely unsatisfactory, and that consequently, the majority of parameters must still be manually optimised.

Diphone Concatenation Synthesiser

Most current commercial applications use various concatenation techniques, i.e. techniques where segments of speech are tied together to form a complete speech chain. These techniques require a fair bit of manual preparation of the appropriate speech segments, but once the segment inventory is constituted, only moderate computational power is needed to chain the segments into an acceptable speech stream.

As indicated above, attempts at building utterances from *phoneme* wave forms have been of limited success, due to coarticulation problems. The use of larger concatenative units has been more successful. Particularly diphones (i.e. excised wave forms from the middle of one phoneme to the middle of the next one) appear to handle coarticulation problems reasonably well. Diphones are concatenated at points were there is a minimum degree of coarticulation, so that transitions at the diphone boundaries must be subjected to a minimum of smoothing.

The development of the linear predictive coding (LPC) technique for speech analysis and re-synthesis has made it possible to store relatively large inventories of high quality speech wave forms in limited space. This model describes a signal sample as a linear combination of the preceding samples. The algorithm calculates model coefficients by minimising the mean square error between the predicted signal and the original signal. These coefficients are recalculated every 5 to 20 ms. About 10 to 16

coefficients are necessary to obtain an acceptable synthesis quality. The system is in fact an all-pole linear filter that simulates the source spectrum and the vocal tract transfer function. The technique has many advantages, such as the automatic analysis of the original signal, fairly easy algorithmic integration, and fidelity to the original sound.

However, among the problems with diphone synthesis remains the danger of major discontinuities occurring at the interface between two halves of a vowel, in cases where dissimilar formant targets are used on the two sides of the interface. In severe cases of formant discontinuity, this produces a bi-vocalic sound quality and an audible discontinuity at the diphone boundary. Also, the model is not well-adapted to nasal and fricative simulation, because the vocal tract model contains only poles, while nasal and fricative spectra involve prominent zeroes, as well as poles.

Synthesis by Vocal Tract Simulation

The preceding techniques attempt, by simple computational techniques, to (re-)generate a signal that is perceptually optimal. They are intended to produce a wave whose spectrum is as close as possible to the real speech signal without simulating any aspect of the human vocal tract. As indicated, such techniques simply attempt to *limit* the negative effects of the coarticulation phenomenon.

A number of relatively recent vocal tract simulation techniques attempt to deal directly with the coarticulation problem by simulating the physical behaviour of the speech production apparatus (Scully, 1990; Maeda, 1990). Such an *articulatory model*;reconstitutes the shape of the vocal tract as a function of the position of the phonatory organs (lips, jaw, tongue, velum). The signal is calculated by a mathematical simulation of the air flow through the vocal tract. The control parameters of such a synthesiser are: sub-glottal pressure, vocal cord tension, and the relative position of the different articulatory organs.

This technique might seem seductive: A more physical model should be easier to control than a functional one. Indeed, in laboratory use, this approach has produced some good preliminary results. However, many problems remain to be solved. For example, articulatory data obtained by cineradiographic recordings are relatively imprecise, and do not produce a complete inventory of articulatory configurations to be used in synthesis. On the other hand, the vocal cord source is difficult to model for vowels, and the source is even more difficult to generate in the case of stops and obstruents. Currently, this kind of synthesiser is therefore still reserved for fundamental research.

A Classification of Synthesis Techniques

It is interesting to consider the classification presented in the preceding paragraphs in terms of a number of operational parameters (see Figure 1):

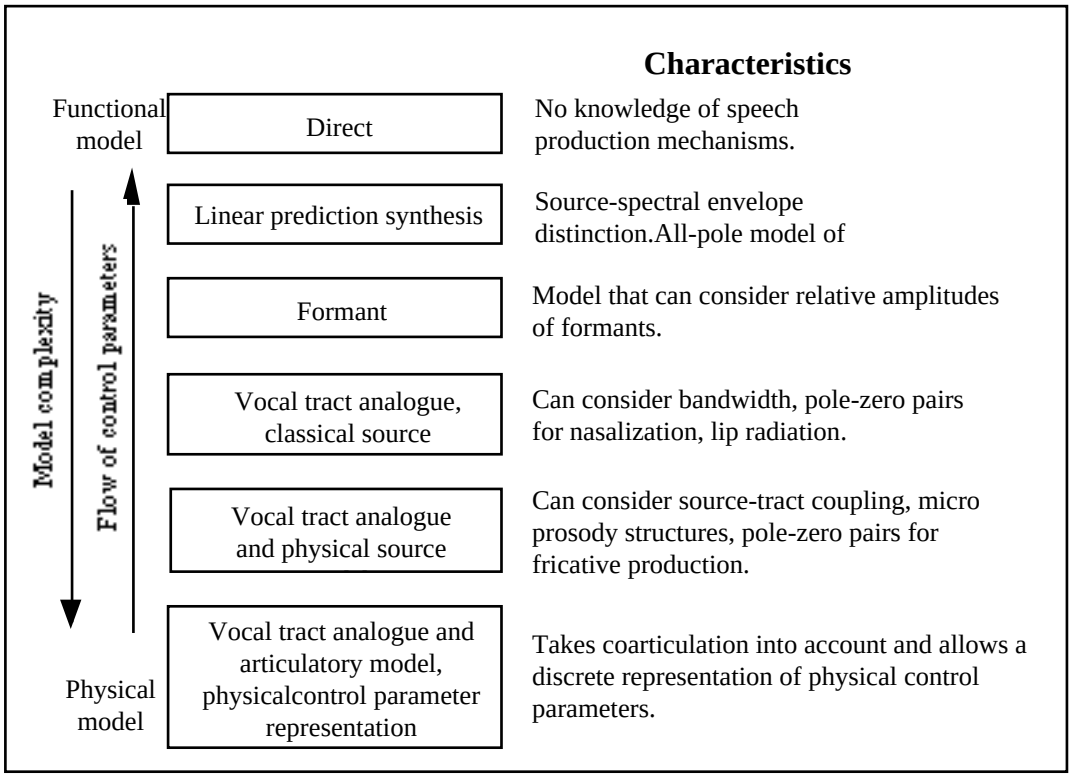

Figure 1. Classification of different synthesis techniques as a function of the model complexity and the flow of the control parameters.

— *decreasing information flow* required to generate the control parameters for synthesis. In direct synthesis, a new command is required each fundamental period cycle, while in the case of the articulatory model, a new command is generated only at the beginning of each articulatory gesture,

— *increasing model complexity* that integrates a progressive knowledge of the production mechanism,

— a progressive distinction between *functional models* that simulate speech by a set of global parameters, and *physical models* that simulate the details of the physical process.

Source-Filter Model

Since the most prominent synthesis techniques in use today are based on the *source-filter* concept, it is interesting to consider in some detail the

different aspects of this theory. The source-filter theory states that the vocal tract can be modelled as a linear filter that varies over time. The filter (i.e. a set of resonators) is excited by a source, which can be either a simulation of vocal cord vibration for voicing, or a noise that simulates a constriction somewhere in the vocal tract. The sound wave is created in the vocal tract, then radiates trough the lips (Figure 2).

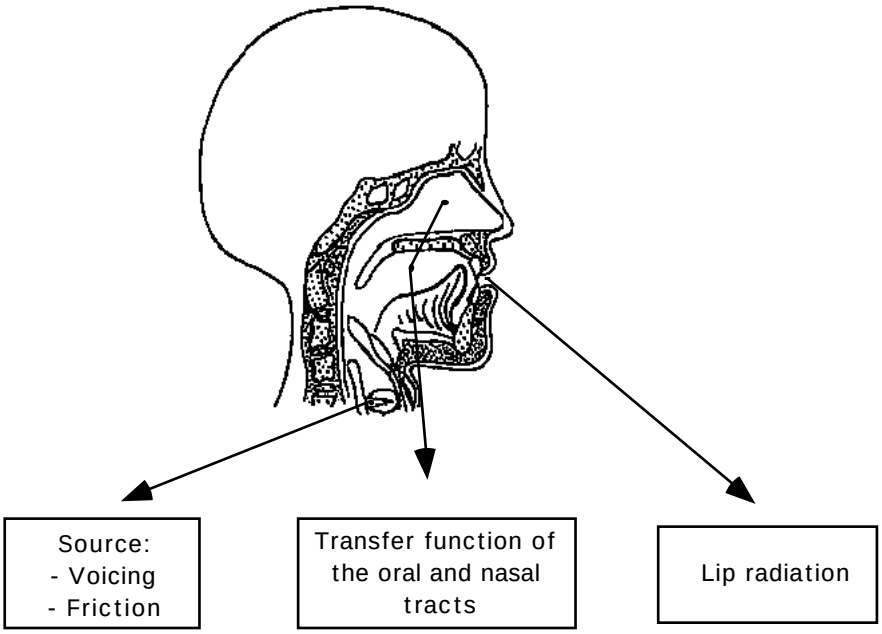

Figure 2. Concept of the source-filter model. The model is divided into tree separate parts, namely the source (voicing and friction), the filter (implementing the transfer function of oral and nasal tract), and lip radiation.

In this model, there is no interaction between the source and the filter, other than that the fact that the filter imposes its resonant characteristics on the source. Hence, the individual acoustic properties of the source and the filter can be separately simulated. The vocal tract filter can be modelled as an acoustic tube with a varying cross-sectional area formed by the pharynx, the oral cavity, the nasal cavity, and the lips. The resonance effects observed in this tube are in turn simulated by a linear filter.

Source Modelling

Let us consider each part of this model in detail.

Speech sounds may be divided into those produced with a periodic vibration of the vocal cords (voiced sounds), and those generated without vocal-cord vibrations, but with plosive or friction noise (voiceless sounds). For this reason, two excitation sources are needed for synthesis:

(i) a source producing a quasi-periodic wave, the voicing source, and

(ii) a noise generator, the friction source.

The Voicing Source

The excitation for voiced sounds provided by the vocal fold vibration can be represented by the general block diagram of Figure 3.

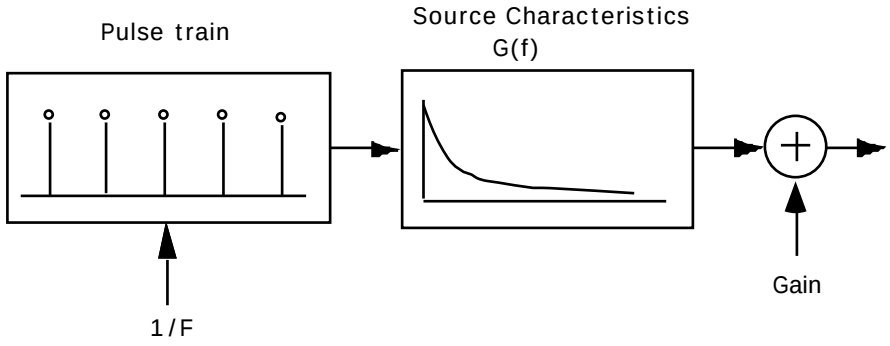

Figure 3. Block diagram of a basic voicing source model. It simulates the glottal wave by filtering an impulse train. A gain control allows the amplitude setting for voicing.

The model is composed of an impulse train generator that produces pulses at the rate of one per fundamental period. This impulse excitation simulates the generation of acoustic energy at the instant of the opening of the vocal cords. This signal then drives a linear filter whose frequency response G(f) approximates the glottal wave form. The function G(f) must be chosen so that it approximates accurately the spectrum of the source. At the end, a gain control device allows the adjustment of the voicing amplitude.

The *source spectral characteristics* are directly related to the type of vocal cord activity that is modelled (Klatt and Klatt, 1990). As shown in Figure 4.a, different maximal glottal openings (as in laryngealised, modal and breathy voice) result in various wave forms. The temporal characteristics of each signal are shown in Figure 4.b, and Figure 4.c shows their respective frequency characteristics. It can be seen that the transfer function has low-pass characteristics. The spectral slope can vary between approximately -12 dB/octave and -24 dB/octave. This value is directly related to the duration of glottal opening. For instance, spectral roll-off is smooth in the case of a short opening duration.

Different models have been proposed for characterising the *transfer function* G(f). The simplest consists of a low-pass filter whose spectral slope can be varied (Klatt, 1980). In the case of the model presented in Figure 3, the impulse train is simply sent through such a filter to obtain a good approximation of the glottal wave form.

Some more recent models represent G(f) with a more accurate mathematical function (Titze, 1989; Fant and Liljencrants, 1985) or a mechanical simulation of the vocal cord vibration (Ishizaka and Flanagan, 1972; Flanagan *et al.*, 1975). These models can introduce some irregularities into the fundamental period cycle, which adds more naturalness to the voice (Klatt, 1987).

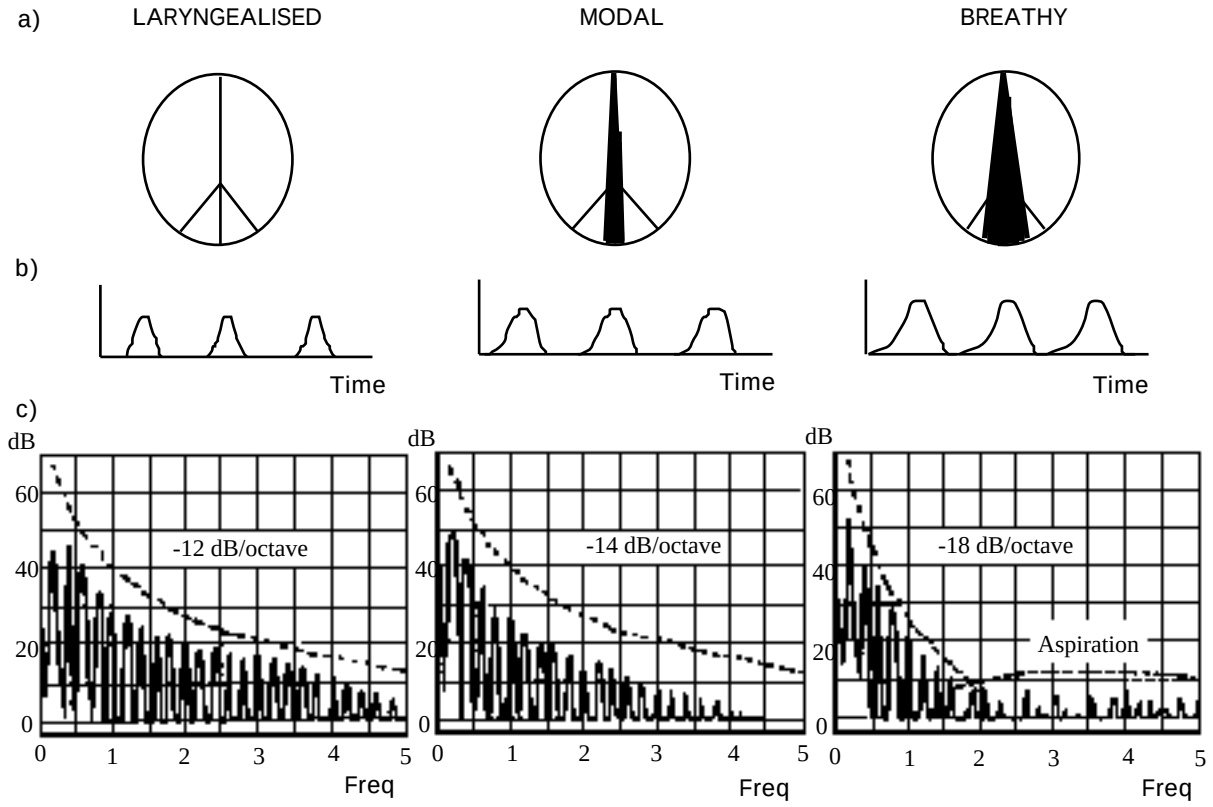

Figure 4. Various glottal source characteristics for laryngealised, modal and breathy vowels. a) Schematic glottal opening for different voice qualities. b) Glottal volume velocity waveform. An increasing opening at the arythenoids is characterised by a longer open period and a less abrupt closure event. c) Source spectra. A progressive opening of the vocal folds is characterised by an increased spectral slope. In case of breathy voicing, the weaker high-frequency harmonics are replaced by aspiration noise (Klatt and Klatt, 1990).

Friction Source

Voiceless sounds are generated when the vocal cords are in a non-vibrating mode and are held open. This allows the air to flow unobstructed through the vocal tract, where it is inhibited at some point to create friction or a plosion. This phenomenon is due to a pressure drop across a constriction formed in the vocal tract, where the flow of air becomes turbulent.

The usual model for the friction source consists of a pseudo-random white noise generator and a gain parameter to allow an adjustment of

friction amplitude. Nevertheless this model is not very accurate, since the actual friction source is not located in the larynx as the model states, but at the location of the main constriction where the airflow becomes turbulent (Badin, 1989). The exact mechanisms involved in the creation of this acoustically complex phenomenon are still under discussion, so that a fully satisfying friction model for synthesis purposes is still some time away.

Vocal Tract Modelling

The source-filter modelling of *vocal tract behaviour* is intended to simulate the formant characteristics of the various speech sounds. It is an efficient approach allowing precise simulation of almost all phonemes. In the initial model that we will discuss here, the resonance characteristics depend only on their respective configuration and are independent of the source. In other words, it is assumed that no source-tract interaction occurs. Thus this supposition is not entirely realistic, because a non-linear coupling exists between the glottal source and the first vocal tract resonance modes (Fant, 1986; Stevens and Bickley, 1986).

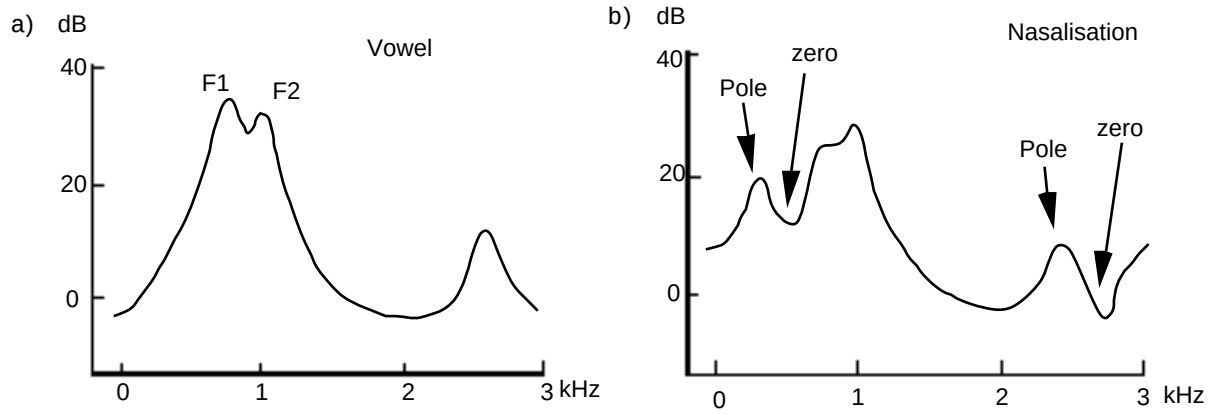

Figure 5. Example of the frequency spectrum of a vowel (a) and a nasalised vowel (b). Oral sounds, such as vowels, can be simulated with an all-pole system (formants), whereas nasalisation implies a pole and zero system (formants and anti-formants).

For oral sounds, such as vowels, the acoustic wave form generated by the source is propagated through the vocal tract. As the acoustic wave form moves through the vocal tract, some of its frequencies are augmented, whereas some others are depressed. Which of the frequencies undergo enhancement depends on the particular configuration of the speech organs, that is, on the position of the jaw, the tongue and the lips. The frequencies occurring at the main resonant areas of the vocal tract for

a given sound are the formants. These resonances may be assumed to behave as a resonating all-pole filter, and can thus be modelled with a set of *poles*. These poles are identified as “peaks” in the frequency spectrum (Figure 5.a).

When the nasal cavities are coupled with the vocal tract, a certain number of anti-resonances (anti-formants) also appear in the spectrum, as a result of the sound-damping properties of the nasal cavity. Those are identified as “troughs” or *zeros* in the transfer function. Hence for nasals and nasalised vowels, the transfer function has to contain both poles and zeroes (Figure 5.b). It should be mentioned that zeroes also appear in the complex articulation of fricatives and stops.

It has been shown that a good approximation can be obtained by modelling each formant with an electrical resonator (Fant 1960). The latter is a band-pass filter (pole filter) whose characteristics are shown in Figure 6.a. Two parameters may be specified, the resonance frequency F and the bandwidth BW. An anti-formant can be modelled with a band-stop filter (zero filter) having the inverse characteristics. Figure 6.b shows the generation of the first four formants as a superposition of resonators at different frequencies.

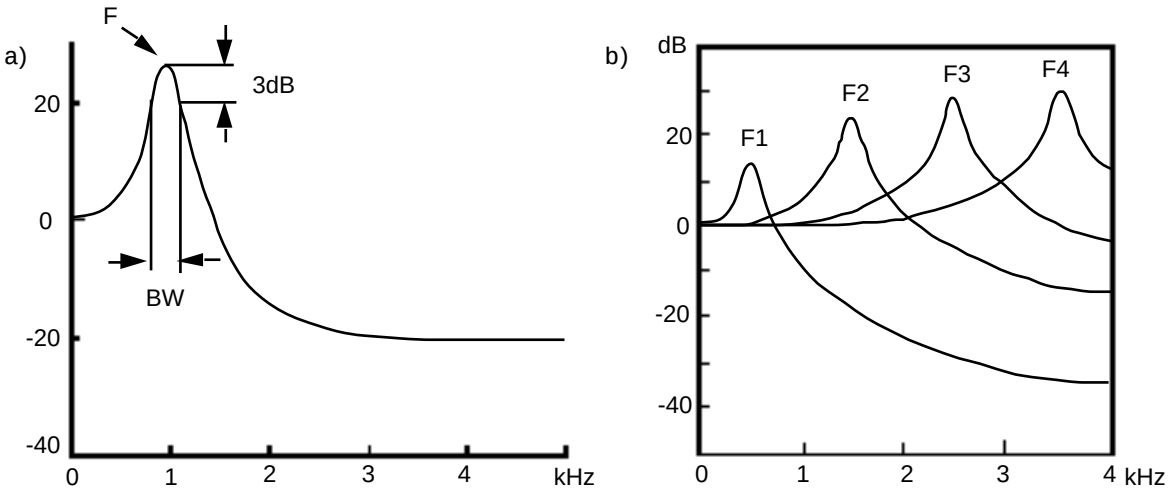

Figure 6. Transfer function of the formant resonators. a) Spectrum of a single formant resonator. The resonator is a second order low-pass filter characterised by its centre frequency F and its bandwidth BW. b) Contribution of each formant to the amplitude spectrum.

The transfer function of the entire tract is obtained by connecting these different circuits according to two possible combinations (Figure 7). In the first configuration, the resonators are combined in parallel. In practice, this corresponds to the successive *addition* of each filter’s transfer function. Each resonator is preceded by a gain which adjusts the relative amplitude of the given spectral peak. In the second case, synthesis operates by connecting the resonators in cascade, a process which results in the *multiplication* of the spectrum by each of the successive transfer functions.

There are theoretical premises for these distinctions. A parallel formant synthesiser allows for the direct control of each formant amplitude and sums the outputs of the simultaneously excited formant resonators. While not an accurate acoustic imitation of vocal tract behaviour in speech, parallel synthesisers are better adapted at producing consonants than vowels (Holmes, 1983). The serial connection of the formant resonators adds the effect of each higher resonance to the final output, and thus produces a direct replica of the total formant energy distribution, which corresponds quite well to the natural resonance mode of the vocal tract. This approach constitutes a fairly faithful imitation of vocal tract behaviour, and as a result, serial synthesisers are particularly good for synthesising vowel sounds.

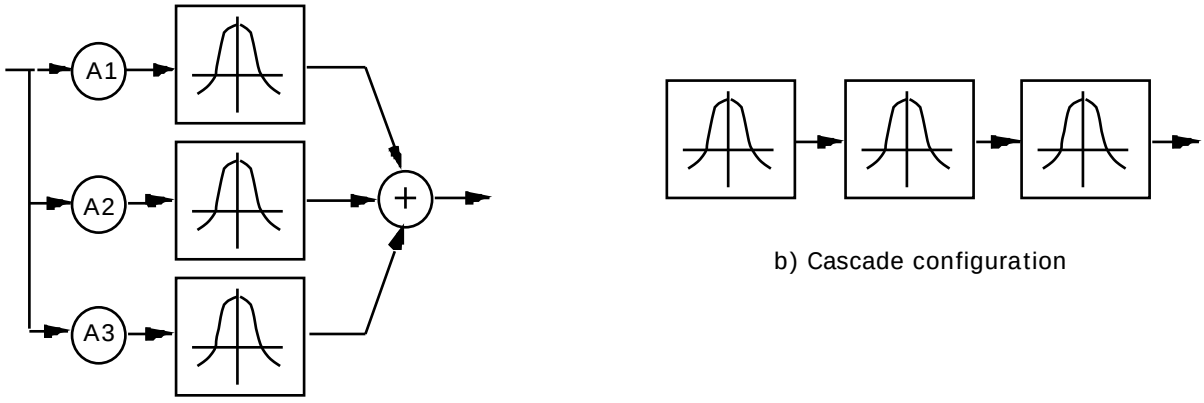

Figure 7. Two ways of combining formant resonators. a) Parallel configuration. b) Cascade configuration.

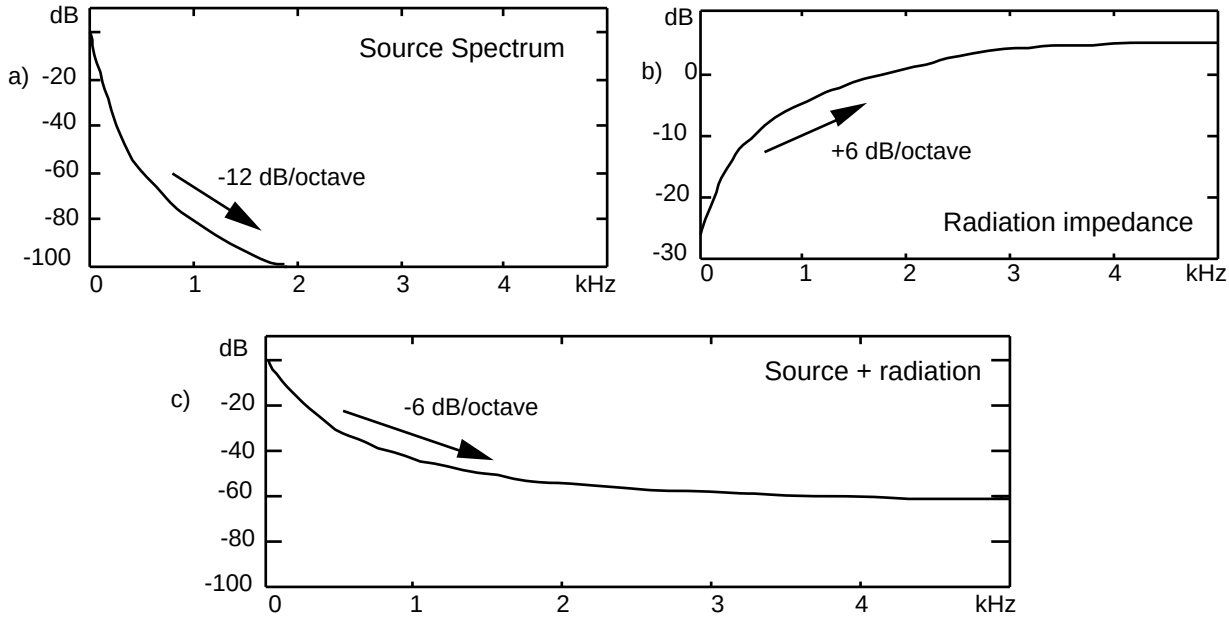

Figure 8. Spectral characteristics of the voicing source and lip radiation. a) Source spectrum with its approximate -12 dB roll off. b) Spectrum of the radiation characteristic. c) Combined effect of the source spectrum and radiation characteristic. The tilts caused by the combined -12 dB/octave roll-off in the glottal airflow and the +6 dB/octave lift of the radiation effect cause the voiced speech spectrum to fall off at an approximate rate of -6 dB/octave.

Lip Radiation

At the mouth opening, the acoustic wave ceases to be constrained in its propagation. This modification of the propagation medium leads to reaction forces that the wave has to counter. This is identified by the modification of the radiation impedance, whose characteristics vary with the wave frequency. The frequency spectrum of a radiated speech sound is tilted upward by approximately +6 dB/octave (Figure 8.b), since high frequency components are better transmitted through the opening than their low frequency counterparts. In synthesis practice, this can be simulated by the application of a high-pass filter.

It has to be noted that the tilts caused by the combined -12 dB/octave roll-off in the glottal airflow and the +6 dB/octave lift of the radiation effect cause the voiced speech spectrum to fall off at an approximate rate of -6 dB/octave (Figure 8.c).

Example of a Formant Synthesiser

To conclude this chapter, we shall outline a prominent implementation of the electrical analogue synthesis technique, i.e. the synthesiser proposed in 1979 by Dennis Klatt (Klatt 1980). This is a direct application of the source-filter theory, implemented by a computer simulation of an electrical structure, consisting of resonators combined in cascade or parallel. The electrical analogue has its historical roots in the first synthesisers that were built of discrete electrical elements. However nowadays, numerical systems are generally used to simulate the operation of these elements.

The Klatt implementation is a cascade/parallel synthesiser, which allows a choice between the formant resonator configurations according to the type of sound to be produced, and which permits the simulation of male and female voices. Its block diagram is shown in Figure 9. A set of 40 parameters determine the output wave. Their abbreviations are explained in Figure 10. Thirty-four of these can be varied dynamically (represented by the symbol “V”). The constant parameters in Figure 10 (symbol “C”) control the general configuration of the synthesiser.

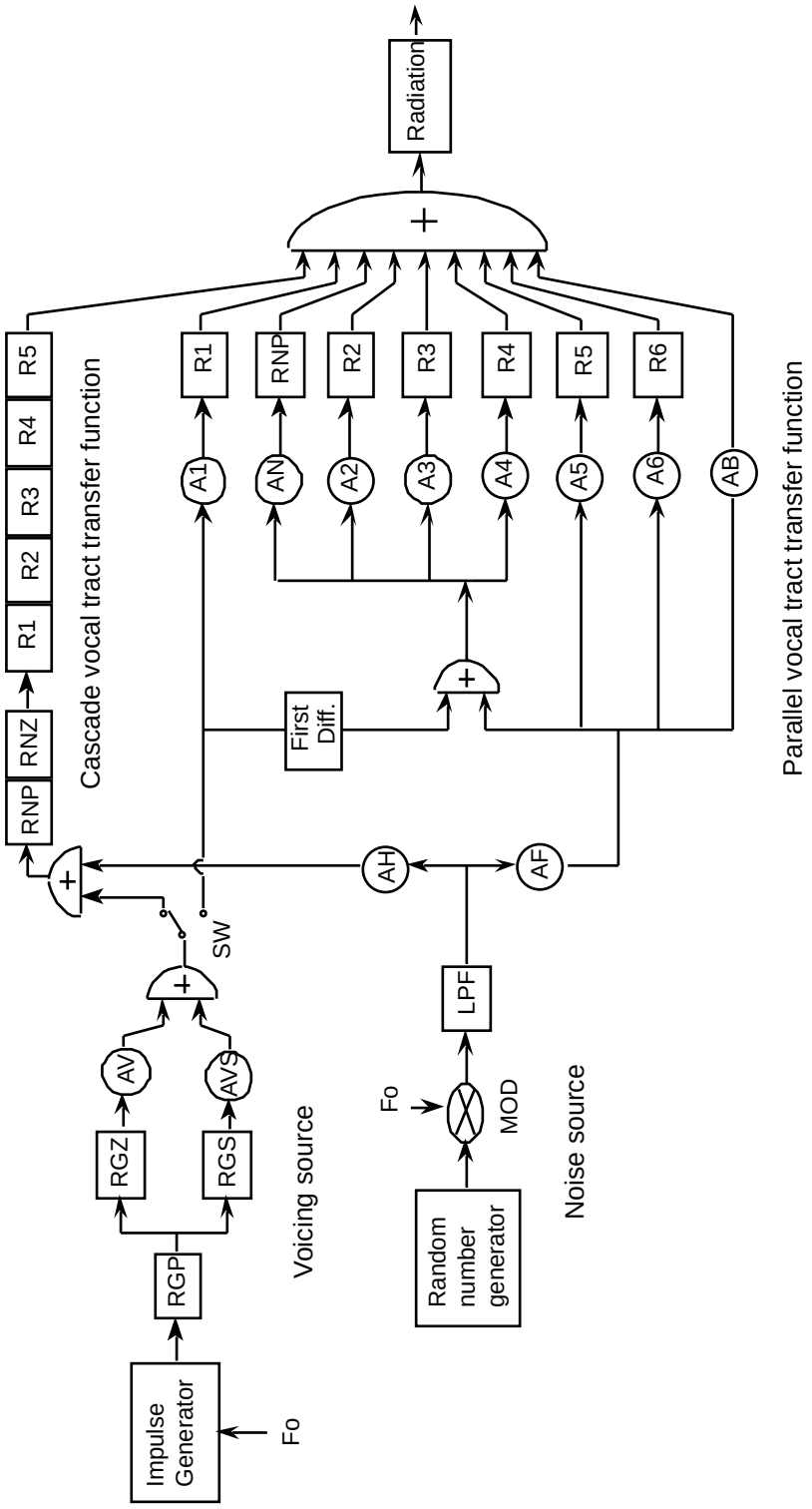

Figure 9. Block diagram of the Klatt 79 synthesiser (Klatt, 1980).

Sources

The synthesiser implements two voicing sources. The first is characterised by a -12 dB/octave spectral slope and is primarily destined for vowel synthesis. A low-pass filter (RGP) performs this function. An anti-

resonator (RGZ) optionally modifies spectral details of the source to some degree. The second voicing source is a quasi-sinusoidal source required for voiced fricatives and voice bars[2]. Its spectral slope (-24 dB/oct.) is obtained with a second low-pass filter (RGS). The switch SW sends both voicing sources either to the cascade (Figure 9, top), or the parallel configuration (Figure 9, bottom).

Symbol	C/V	Min.	Max.	Name
DU	C	30	5000	Duration of the utterance (ms)
NWS	C	1	20	Update interval for parameter reset (ms)
SR	C	5000	20000	Output sampling rate (Hz)
NF	C	1	6	Number of formants in cascade branch
SW	C	0	1	0=Cascade, 1=Parallel tract excitation by AV
G0	C	0	80	Overall gain scale factor (dB)
F0	V	0	500	Fundamental frequency (Hz)
AV	V	0	80	Amplitude of voicing (dB)
AVS	V	0	80	Amplitude of quasi-sinusoidal voicing (dB)
FGP	V	0	600	Frequency of glottal resonator "RGP"
BGP	V	50	2000	Bandwidth of glottal resonator "RGP"
FGZ	V	0	5000	Frequency of glottal anti-resonator "RGZ"
BGZ	V	100	9000	Bandwidth of glottal anti-resonator "RGZ"
BGS	V	100	1000	Bandwidth of glottal resonator "RGS"
AH	V	0	80	Amplitude of aspiration (dB)
AF	V	0	80	Amplitude of frication (dB)
F1	V	180	1300	Frequency of 1st formant (Hz)
B1	V	30	1000	Bandwidth of 1st formant (Hz)
F2	V	550	3000	Frequency of 2nd formant (Hz)
B2	V	40	1000	Bandwidth of 2nd formant (Hz)
F3	V	1200	4800	Frequency of 3rd formant (Hz)
B3	V	60	1000	Bandwidth of 3rd formant (Hz)
F4	V	2400	4990	Frequency of 4th formant (Hz)
B4	V	100	1000	Bandwidth of 4th formant (Hz)
F5	V	3000	6000	Frequency of 5th formant (Hz)
B5	V	100	1500	Bandwidth of 5th formant (Hz)
F6	V	4000	6500	Frequency of 6th formant (Hz)
B6	V	100	4000	Bandwidth of 6th formant (Hz)
FNP	V	180	700	Frequency of nasal pole (Hz)
BNP	V	40	1000	Bandwidth of nasal pole (Hz)
FNZ	V	180	800	Frequency of nasal zero (Hz)
BNZ	V	40	1000	Bandwidth of nasal zero (Hz)
AN	V	0	80	Amplitude of nasal formant (dB)
A1	V	0	80	Amplitude of 1st formant (dB)
A2	V	0	80	Amplitude of 2nd formant (dB)
A3	V	0	80	Amplitude of 3rd formant (dB)
A4	V	0	80	Amplitude of 4th formant (dB)
A5	V	0	80	Amplitude of 5th formant (dB)
A6	V	0	80	Amplitude of 6th formant (dB)
AB	V	0	80	Amplitude of bypass path (dB)

Figure 10. Control parameters of the Klatt ’79 synthesiser. A set of 40 parameters determine the output wave. Thirty-four can be varied dynamically (symbol “V”). The constant parameters (symbol “C”) control the general configuration of the synthesiser.

[2] I.e. the low-frequency energy bands resulting in normal speech from the voice fundamental frequency.

A random noise generator is also available and furnishes the equivalent of a friction source controlled by the gain AF and an aspiration source, with gain AH, that is mixed with the voicing source. As the noise source spectrum must be approximately flat, a low pass filter (LPF) cancels the effect of lip radiation (represented in the last block in Figure 9). In the case of voiced sounds, the turbulence noise is modulated by the vocal fold vibration. This amplitude modulation (MOD) is controlled by the fundamental frequency. One may note that the friction source is only sent through the vocal tract in its parallel configuration.

Vocal Tract Simulation

The vocal tract, in its cascade configuration, is realised with five resonators (R1…R5), whose central frequency and bandwidth can be individually adjusted. The addition of a supplemental resonator (RNP) and anti-resonator (RNZ) allows synthesis of nasal sounds. The parallel configuration contains seven formant resonators (R1…R6, RNP), each having an individual gain control (A1...A6, AN). Moreover a bypass connection, containing only a volume control (AB), allows simulation of sounds that do not have prominent peaks in their spectrum.

Developments Since 1979

The synthesiser presented in this paragraph is not the latest version of Klatt's formant synthesiser. In the past 10 years, substantial improvements have been made to allow a better synthesis of male and female voices.

Perhaps the most significant improvements are related to the glottal source. Parametric source models, such as proposed by Fant and Liljencrants (Fant and Liljencrants, 1985), or the one proposed by Klatt (Klatt and Klatt, 1990), allow a better waveform shape generation than the impulse source presented earlier. To match the voice quality of different speakers, parameters such as the glottal opening time during a voicing period, and the velocity of the glottal closure, may be adjusted. Within an utterance by a given speaker, these parameters can be varied as active laryngeal adjustments are made to produce voiceless obstruents consonants or prosodic changes within phrases and sentences. The parameters can also be modified, as the laryngeal state reacts passively to the manipulation of constrictions in the airway, for example during voiced obstruents and or sonorant consonants produced with a narrow constriction. In addition, to improve naturalness of the synthesised

speech, some glottal pulse timing irregularities, such as jitter and diplophonia, can now be added.

Proper adjustment of these parameters permits the generation of a glottal source with a spectrum that is a good approximation to the spectrum of glottal source for almost any male or female speaker. Nevertheless, a deeper understanding of the acoustic manifestations of variations in voice qualities and during the production of various types of sounds is needed. An important work that goes in this direction has been performed by Gobl (1988), who studied variations of the voice source in connected speech by means of inverse filtering and waveform parametrisation. This study showed significant changes in the glottal pulse shape found at the onset and at the termination of the voice source, and also at many of the boundaries between vowels and consonants. Re-synthesis of utterances using parameters from an acoustical analysis showed good results.

It has been stated above that the source-filter model is a linear system in which source and vocal tract interactions are neglected, on the assumption that the volume velocity waveform depends very little on the shape or impedance of the vocal tract. It is also known that this assumption is not entirely correct, since the vocal fold impedance can vary, due to different glottal openings, and because of constrictions that have a “retro-effect” on vocal fold vibration. The presumed relationship between glottal area and glottal flow is perturbed by standing wave-pressure fluctuations in the pharynx, which invalidate the assumed transglottal pressure over a cycle.

Some of these non-linear coupling effects can be simulated in Klatt’s new synthesiser version (Klatt and Klatt, 1980). In the cascade branch, an additional pole-zero pair is inserted, which allows simulation of acoustical coupling to the trachea when the glottal opening is sufficiently large. In addition, these supplemental resonators can be used for better nasal sounds synthesis. Especially nasalised vowels can be simulated with a better accuracy. The time-varying glottal impedance affects the vocal tract transfer function primarily by causing increased losses at low frequencies when the glottis is open. It has also been observed that the first formant frequency may increase during the open phase of the glottal cycle. A method for synchronously changing first formant frequency and bandwidth pitch is thus provided.

Another improvement made by Cheng and Guérin (1987) to the parallel configuration consists in introducing a pole-zero pair in cascade with the first formant resonator. They have shown that the parallel configuration allows a better adjustment of the overall spectrum details for the synthesis of nasalised vowels. The introduction of a zero is necessary to provide the nasal percept.

The generation of speech with a formant synthesiser having a large set of control parameters requires that quantitative data and explicit models

be developed in two areas of phonetics. One is concerned with constraints that the articulatory and aerodynamic systems impose on the sound. The other area involves the temporal control of the articulatory processes that determine the timing of the control parameters. Developing mapping relations requires that theories and models of glottal vibration and vocal-tract acoustics be refined, for example, by the estimation of the distribution of turbulence noise with vocal tract constrictions, by the determination of the time course of onsets and offsets of vocal fold vibration for voiced consonants, and by the modelling of acoustic losses with consonantal constrictions. Refinements in our understanding of articulatory control processes highlight the need for several types of data and models. Quantitative data must be obtained on rates of release and closure of articulators that form the primary consonantal constrictions for stops and fricatives. Furthermore, it is necessary to determine how articulatory parameters that are not directly involved in forming the consonantal constriction are timed in relation to the primary articulators.

Conclusion

In this chapter, a brief review of the principal speech synthesis methods has been presented. Diphone concatenation systems are currently the most widely used synthesis technique. Various commercial systems are available, which produce a fairly good synthesis.

The limitations of these systems are mainly related to the excessive size of diphone data bases that are required to deal with coarticulation problems. Relatively large speech segments (e.g. demi-syllables or syllables) are needed to handle coarticulation more or less satisfactorily. To attain even more satisfactory diphone speech, even larger segments may be necessary, which in turn would lead to even larger data bases.

One alternative is formant synthesis which operates on space-efficient, parametrised input. Given a coarticulatorily motivated string of input parameters, formant synthesisers solve the coarticulation problem quite elegantly. However, current problems concern the precise definition of, and the exact balance between these input parameters, which are quite delicate. However, recent work has shown that automatically generated formant synthesis can produce speech virtually indistinguishable from model utterances. The main disadvantage of these systems remains the considerable computational load which currently allows real-time synthesis only on powerful computers.

References

Atal, B.S., & Hanauer, S.L. (1971). Speech synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America, 50,* 637-655.

Badin, P. (1989). Acoustics of voiceless fricatives: Production theory and data, *Speech Transmission Laboratory Quatrerly Progress and Status Report, 3,* 33-55.

Calliope (1989). *La parole et son traitement automatique*. Masson.

Cheng, Y.M., & Guérin, B. (1987). Nasal vowel study: Formant structure, perceptual evaluation and neural representation in a model of the peripherial auditory system. *Bulletin de la Communication Parlée, 1,* 91-132.

Cooper, F.S., Liberman, A.M., & Borst, J.M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences of the United States of America, 37,* 318-325.

Dudley, H. (1939). The vocoder. *Bell Laboratories Record, 17,* 122-126.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fant, G. (1986). Glottal flow: Models and interaction, *Journal of Phonetics, 14,* 393-399.

Fant, G., & Liljencrants, J. (1985). A four parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress and Status Report, 2,* 18-24.

Flanagan, J.L, Ishizaka, K., & Shipley, K.L. (1975). Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *Bell Systems Technical Journal, 54,* 485-506.

Gobl, C. (1988). Voice source dynamics in connected speech. *Speech Transmission Laboratory Quarterly Progress and Status Report, 1,* 123-159.

Gold, B., & Rabiner, L.R. (1968). Analysis of digital and analog formant synthesizers. *IEEE Transactions on Audio and Electroacoustics 16, 1,* 81-94.

Holmes, J.N. (1983). Formant synthesizers: Cascade or parallel. *Speech Communication, 2,* 251-273.

Ishizaka, K., & Flanagan, J.L. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Systems Technical Journal, 51,* 1233-1268.

Klatt, D.H. (1980). Software for a cascade / parallel formant synthesizer. *Journal of the Acoustical Society of America, 67,* 971-995.

Klatt, D.H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America, 82,* 737-793.

Klatt, D.H., & Klatt, L.C. (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87,* 820-857.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 131-149). Amsterdam: Kluwer.

O'Shaughnessy, D., Barbeau, L., Bernardi, D., & Archambault, D. (1988). Diphone speech synthesis, *Speech Communication, 7,* 55-65.

Scully, C. (1990). Articulatory synthesis. In W.J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 151-186). Amsterdam: Kluwer.

Stevens, K.N., & Bickley, C.A. (1986). Effect of vocal tract constriction on the glottal source: Experimental and modeling studies. *Journal of Phonetics, 14,* 373-382.

Titze, I.R. (1989). A four parameter model of the glottis and vocal fold contact area. *Speech Communication, 8,* 191-201.

ANNEX: Terminal Analog Synthesiser

The terminal analog method (or formant synthesiser) simulates the speech production mechanism using an electrical structure consisting of the connection of several resonance (formant) and anti-resonance (anti-formants) circuits.

The complex frequency characteristics (Laplace transformation) of a resonance;(pole) circuit can be represented as (Gold and Rabiner, 1968):

$$H(s) = \frac{s_n \; s_n^*}{(s - s_n)(s - s_n^*)} \tag{1}$$

where

$$s = -\sigma + j\omega$$

$$s_n = -\sigma_n + j\omega_n \quad \text{and} \quad s_n^* = -\sigma_n - j\omega_n \tag{2}$$

Digital simulation of this circuit can be represented with its z-transformation:

$$H(z) = K_p \left(\frac{A z^{-1}}{1 + C z^{-1} + B z^{-2}} \right) \tag{3}$$

where:

$$K_p = \frac{\sigma_n^2 + \omega_n^2}{\omega_n} \qquad A = e^{-\sigma_n T} \sin(\omega_n T)$$

$$B = e^{-2\sigma_n T} \quad \text{and} \quad C = 2e^{-\sigma_n T} \cos(\omega_n T) \tag{4}$$

T is the sampling period. These equations imply that the digital simulation circuit can be represented as shown in Figure A.1(a). When the resonance frequency $f_n = w_n / 2\pi$ [Hz] and bandwidth $b_n = s_n / \pi$ [Hz] are given, the circuit parameters can be obtained. The anti-resonance (zero) circuit indicated in Figure A.1(b) can easily be obtained from the resonance circuit, based on the inverse circuit relationship. Here, $K_z = \omega_n / \sigma_n^2 + \omega_n^2$.

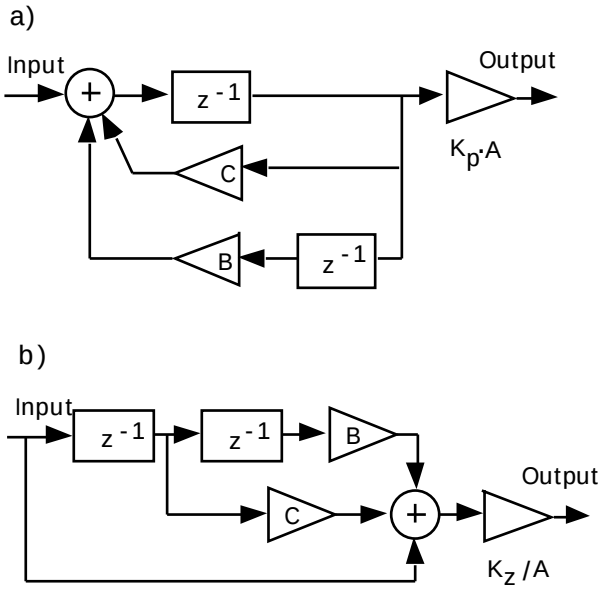

Figure A.1. Digital network implementing a second order resonator. a) Band-pass filter. b) Stop-band filter.

The complete formant network H(z) is obtained by connecting N formant or anti-formant resonators $H_n(z)$ in cascade:

$$H(z) = \prod_{n=1}^{N} H_n(z) \qquad (5)$$

or in parallel:

$$H(z) = \sum_{n=1}^{N} A_n \; H_n(z) \qquad (6)$$

The transfer function of the entire synthesiser can thus be given by:

$$F(z) = G(z) \left(1 - z^{-1}\right) H(z) \qquad (7)$$

Where G(z) is a suitable source filter and H(z) is the formant network. In the early Klatt formant synthesizer version (Klatt, 1980) the voicing source transfer function is approximated with one or two second order resonators (3), where the resonance frequency f_n is set to zero to obtain a low-pass characteristic. The transfer function $(1-z^{-1})$ approximates the lip radiation characteristics.