

## A NEW SYNTHESIS MODEL FOR AN ALLOPHONE BASED TEXT-TO-SPEECH SYSTEM

L. BOVES (\*), J. KERKHOFF (\*), H. LOMAN (\*)

### INTRODUCTION

Although electronic speech synthesis by now has a tradition of several decades, there is still no agreement on the most preferable structure for a speech synthesizer. In this paper we will compare several structures that have been used by workers in the field. As these all appear to have some drawbacks, we will propose an alternative structure that should solve at least some of the problems.

The single most important axiom underlying our work is the opinion that the development of synthesis rules will be made much easier and less time consuming if optimal use can be made of existing phonetic knowledge. This knowledge happens to be formulated either in terms of articulatory postures and movements or in terms of formant patterns. Taking recourse to the acoustic theory of speech production [1,2] it is not too difficult to translate articulatory data into formant patterns. The transformation of formant patterns into articulatory configurations is more difficult; also, the result is not necessarily unique [3]. This is one reason why articulatory synthesis has received much less attention than formant synthesis or terminal analog synthesis. In this paper the discussion will be restricted to terminal analog synthesis, and more specifically, to formant synthesis. The use of linear prediction parameters like reflexion coefficients or Log Area Ratios is not considered because, regardless of their suggestive names, the relation of these parameters to actual vocal tract configurations is, at best, disputable.

### FORMANT SYNTHESIZERS, A CRITICAL APPRAISAL.

Regardless of the detail of their architecture and implementation all formant synthesizers are based on the source-filter theory of speech production. Thus they contain a source part, that represents the activity of the voice source, and a noise source that represents the turbulence caused by air flowing past an obstacle or constriction in the vocal tract, whereas the acoustic filtering operation of the vocal tract is accounted for by a network of resonances (and anti-resonances). The effect of radiation at the lips and nostrils is most often combined with the source section of the synthesizer.

The resonance and anti-resonance networks that represent the vocal tract can be arranged in three different ways: in series, in parallel or partly in series, partly in parallel. Examples of completely serial or parallel connections can be found in [4,5]; serial-parallel architectures are described in [3,6,7,8]. A detailed discussion of the pros and cons of the architectures is given in [5]; therefore, we can confine ourselves to a summary of the most salient aspects.

Synthesizers consisting of a serial connection of resonance circuits are well suited for the synthesis of vowels, but they have great difficulties in producing many consonants. One of their attractive features is that the resonance circuits can be controlled by two parameters, centre frequency and bandwidth. Parallel synthesizers can cope with both vowels and consonants, but at the expense of a considerable increase in complexity. For one thing, the resonator circuits need three parameters for their control, viz. amplitude in addition to frequency and bandwidth. Besides that, special precautions are necessary for maintaining a correct spectral level at very low frequencies. Combined serial-parallel synthesizers share the disadvantage of the increased complexity. This

-----  
(\* Institute of Phonetics, Nijmegen University, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

may not be a real disadvantage, however; The simplicity of control in a strictly serial synthesizer may after all be deceptive in that it does not allow the minimum level of control necessary for adequate speech synthesis. Virtually all existing formant synthesizers have one thing in common, what, in our view, is their fundamental weakness: they are truly terminal analog systems, that attempt to approximate a given spectral envelope, without explicit reference to (and sometimes necessarily explicitly disregarding) the mechanism of speech production. In the next section we will propose a synthesis architecture that overcomes this limitation.

### A SERIAL POLE-ZERO SYNTHESIZER.

As long as the glottal impedance is very high, the velum is raised so as to disconnect the nasal cavity, and the lip opening is appreciable, the vocal tract can be represented by an unbranched non-uniform tube, the acoustic behaviour of which is completely specified by the frequencies and the bandwidths of its resonances. The all-pole characteristic of the vocal tract gets lost as soon as the velum is lowered and the tract is changed into a branched tube, or when the point of excitation shifts away from the glottis to some constriction in the vocal tract. In all these cases the frequency response of the vocal tract contains zeros in addition to poles. The zeros may have two effects: They may introduce sharp dips in the spectrum or they may lower the overall spectral level in a fairly wide frequency band.

Some synthesizers try to account for the anti-resonances in nasal sounds by adding a series connection of a resonance and an anti-resonance. During non-nasal sounds the parameters of the nasal pole-zero pair are chosen so as to cancel each other. In nasal sounds the pole-zero pair is given parameter values that cause an extra spectral peak at a very low frequency and a sharp spectral dip at a slightly higher frequency. From a spectral point of view the synthesis network is more complex during nasal sounds, but this increased complexity can be motivated from an artic-

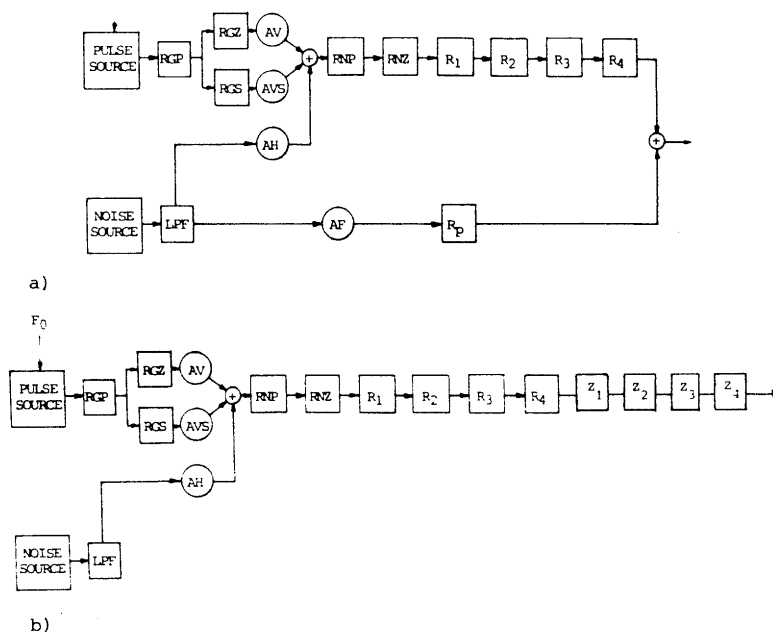


Fig. 1:  
 a) Block diagram of a serial-parallel formant synthesizer with one pole-zero pair.  
 b) Block diagram of a serial formant synthesizer with four additional zeros.

Digital resonators are indicated by the prefix R, anti-resonators, by the prefix Z, amplitude controls by the prefix A.

ulatory point of view. The addition of pole-zero pairs that cancel each other when not needed

cannot be used to solve the problems in approximating the correct spectral envelope for consonants that have their point of excitation somewhere in the vocal tract. In this situation the cavity in front of the point of excitation acts as a resonator, whereas the cavity behind the excitation absorbs acoustic energy at its resonance frequencies and thus acts as an anti-resonator. Because neither cavity is very long, the number of resonances and anti-resonances in the frequency range between 0 and 5 kHz is limited. In fact the spectra of most fricatives and plosive bursts are known to be well represented by about two resonances and two anti-resonances. Thus it is not surprising that some synthesizers contain an extra branch, consisting of a series connection of a small number of number of resonances and anti-resonances, that is parallel to the vowel/nasal branch and that is only excited during the production of several unvoiced consonants (cf. Fig 1a). This argument requires that considerable caution is exercised in switching between the branches in such a way that the auditory continuity of the output signal is maintained. Also, switching between the essentially independent branches cannot easily be motivated from an articulatory point of view. In order to profit from an anti-resonance in a series connection of a resonance and an anti-resonance, the latter must be moved from its cancelling position in the low frequency region of the spectrum. That, however, results, in an extra pole. Although there may be cases that this extra pole too can be used, simply for spectral shaping purposes, somewhere higher in the frequency spectrum, the speed and magnitude of the pole movement is so great in such cases that disturbing noises are introduced in the output signal. An example of this phenomenon can be observed in Fig. 2. The natural speech signal (Fig. 2a) and the synthesized signal (Fig. 2b) resulting from the model in Fig. 1a match reasonably well, due to the position of the nasal pole and zero (3000 Hz and 1800 Hz respectively). Matching both spectra only by means of the parallel resonator in combination with the serial resonators was less succesful, so adding a zero was necessary. The disturbing noise occurs during the transitional phase of the nasal pole (from 240 to 3000 Hz).

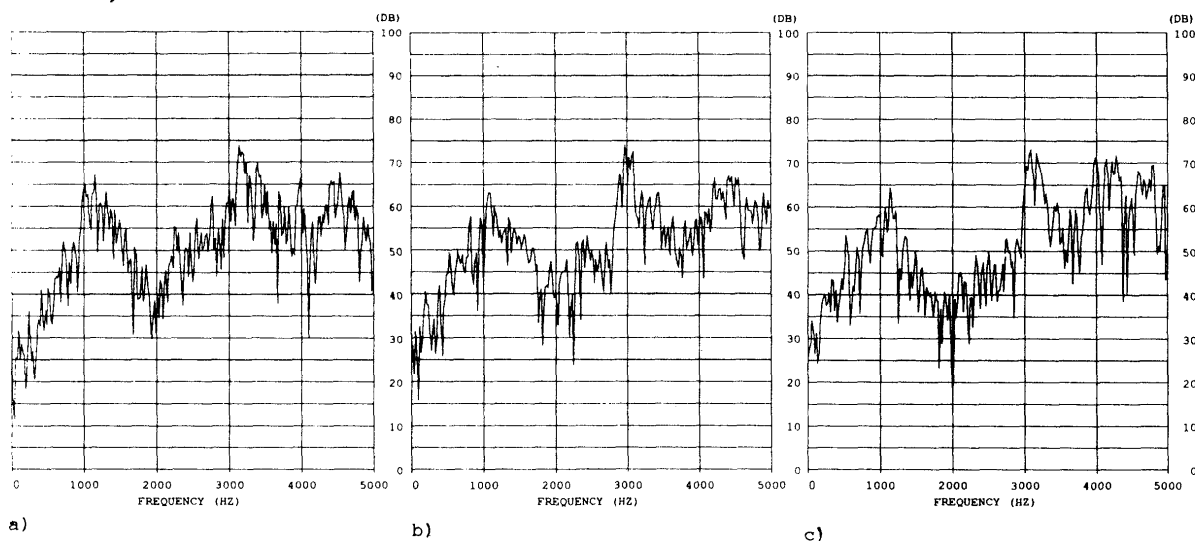


Fig. 2: FFT spectra of 25 ms samples of frication noise.

- a) From natural spoken /x/
- b) From rule-synthesized /x/ by means of the model shown in Fig. 1a.
- c) From rule-synthesized /x/ by means of the model shown in Fig. 1b.

One way for synthesizing fricatives that is compatible with articulatory considerations is to add anti-resonance circuits in series to a serial formant synthesizer. Obviously, these anti-resonances must be canceled during the production of sounds that are well approximated by an all-pole spectrum. The only way left to cancel the zeros is to put them into the origin of the z-plane during the time that they are not needed. If their action is called for, the zeros must be shifted towards the unit circle in the z-plane. Because the distance travelled by the zeros is relatively large, their speed of movement is accordingly high. Now it is well known that fast changes in the parameters

of a resonance circuit cause intrusive noises to occur in the output signal that may even affect the auditory continuity of the signal. Contrary to poles, however, zeros are allowed to move around the z-plane at high speed without destroying the auditory continuity and without causing intrusive sounds to be heard. Zero tracks are even allowed to cross without producing objectionable auditory effects. This has led us to the synthesis structure shown in Fig. 1b. Although this architecture still does not allow a direct mapping of articulatory data to synthesizer control parameters it is a viable approximation to the ideal. Switching between two branches is no longer necessary, all-pole sounds can fully profit from the simple control structure of serial formant resonators, F-patterns during the transition of vowels into consonants maintain their articulatory interpretation and the spectral envelopes of consonants can be realized without the need to tamper with the frequency and/or bandwidth of formants of adjacent vowels in order to prevent poles from moving too fast. Fig. 2a and 2c show spectra from a natural speech signal and from a synthetic signal produced with the new pole-zero serial synthesizer. It can be seen that fricative spectra can be closely matched.

### SOURCE MODELS IN FORMANT SYNTHESIZERS.

An often cited disadvantage of serial formant synthesizers is that they are not able to model spectral slope changes due to changes in the shape of the glottal excitation pulses. One solution to this problem would be to employ the added zeros to serve this purpose, but we consider this as abuse. In our opinion these spectral slope changes should be accounted for at the place where they originate, viz. in the source. Here too, two orthodox and a compromise approach can be conceived. One orthodox approach is to opt for a maximally simple terminal analog realization that just produces the desired waveforms without any reference to physiological reality. An example of this approach is the source in the synthesizer described by Klatt [6] and represented in Fig. 1a by the parameters RGP, RGZ and RGS, which is sufficiently transparent to allow rules for its control to be derived without too much effort. It is, however, not very well suited to model details of the dynamic changes of the voice source characteristics. The alternative orthodox solution would be to implement a physiologically based source model like the one described in [9]. The computational complexity of such a model would, however, be prohibitive in any realistic synthesizer; also, it would be extremely difficult to obtain the data necessary to derive the rules for controlling the parameters of the model. Compromise solutions either specify the glottal waveform by a concatenation of mathematical functions [10,11] or by using a drastically simplified physiological model that still can account for the most important aspects of the behaviour of the human voice source [12]. Work is under way to replace the terminal analog source in our synthesizer with a computationally efficient version of the model proposed by Cranen and Boves [12].

### LIST OF REFERENCES.

- [1] Fant, G., Acoustic theory of speech production (Mouton, The Hague, 1960).
- [2] Flanagan, J.L., Analysis, synthesis and perception of speech (Springer, Berlin, 1972).
- [3] Atal, B.S. et. al., JASA, 1535, (1978).
- [4] Rabiner, L.R., JASA, Vol. 43, 822 (1968).
- [5] Holmes, J.N., Speech Communication, Vol. 2, 251, (1983).
- [6] Klatt, D.H., JASA, Vol. 67, 971, (1980).
- [7] Slis, I.H., Proceed. Inst. of Phonetics Nijmegen, Vol. 2, 83, (1978).
- [8] Kerkhoff, J.P. et. al., Proceed. Inst. of Phonetics, Vol. 10, 75, (1986).
- [9] Titze, I.R. & D.T. Talkin, JASA, Vol. 66, 60, (1979).
- [10] Fant, G., STL-QPSR, nr. 1, 85, (1979).
- [11] Fujisaki, H. & M. Lungqvist, Proceed. ICASSP-87, 637, (Dallas, 1987).
- [12] Cranen, B. & L. Boves, JASA, Vol. 81, 734, (1987).