Proceedings of The Institute of Acoustics

Selection of a formant synthesiser model for Text-to-Speech synthesis

D.A. Sinclair

Speech Research, IBM UK Scientific Centre,
Athelstan House, St Clement Street, WINCHESTER, SO23 9DR

Introduction

The choice of a particular synthesis strategy in a text-to-speech system is influenced by a wide range of conflicting requirements. The synthesiser must be capable of producing natural sounding speech, from a narrow phonetic input, at reasonable computational cost and for a range of speaker qualities and styles. Synthesis by rule using formant or LPC parameters has been the preferred choice as this can give acceptable speech at a reasonable computation cost [1,2,3,4,5]. One major task in using such synthesisers within a text to speech system is the determination of the relationship between the phonetic description of the required speech and the corresponding parameters required to drive the synthesiser. These parameters must be determined by examining a corpus of natural speech. Clearly there are advantages in automating this process as far as possible and in adopting a synthesiser architecture that minimises the substantial effort in determining a full library of parameters. This paper examines some design criteria for a formant based synthesiser and the corresponding parameter measurement methods. The reduction of the parameter data into a set of cosegmentation rules is discussed elsewhere [6].

The all pole model of speech production can be characterised by the following transfer function:

$$H(f) = \left| \prod_{i=1}^{n} \frac{1}{(1 - p_i z^{-1})} \right|_{z = e^{j2\pi fT}} \tag{1}$$

where T is the sampling interval and p_i are the poles of function in the z plane. Such a model is a good representation of the non nasal sonorants.

The cascade formant model of speech production can be characterised by the following transfer function:

$$H(f) = \left| \prod_{i=1}^{m} \frac{1 - \alpha_i - \beta_i}{1 - \alpha_i z^{-1} - \beta_i z^{-2}} \right|_{z = e^{j2\pi fT}} \tag{2}$$

$$r_i = e^{-\pi b_i T} \quad \alpha_i = 2r\cos(2\pi f_i T) \quad \beta_i = -r^2$$

where f_i and b_i are the formant centre frequencies and bandwidths. These parameters are related to the pole positions of the transfer function as:

$$f_i = \tan^{-1}\frac{(p_{Ii}/p_{Ri})}{2\pi T}$$

$$b_i = \frac{\log_e |p_i|}{2\pi T} \qquad (3)$$

where p_{Ri} and p_{Ii} are the real and imaginary parts of the *ith* pole.

There are a number of possible techniques for determining synthesiser parameters from a natural speech utterance [7,8,9]. Here we adopt the following approach:

1. The LPC autocorrelation coefficients are determined from a digitised waveform sampled at 10 KHz, using 14 LPC coefficients, an analysis window width of 192 samples, hamming windowing, and a step size of 100 samples between analysis frames.
2. The roots of the denominator polynomial in (2) are determined using a polynomial zero finding algorithm. These roots correspond to the positions of the poles of the speech transfer function.
3. The frequencies and bandwidths of the poles are determined from (3).
4. Valid formants are deemed to be those with bandwidths of less than 500 Hz and whose centre frequencies are not equal to 0 Hz or half the sampling frequency.
5. Poles which do not satisfy the criterion in (4) above are assumed to be associated with overall spectral shaping of the speech transfer function.

This measurement technique has been applied to a range of vowel and non vowel segments for a male RP speaker and the results are shown in Figure 1.

One major advantage of the formant synthesiser over LPC resynthesis is the possibility of removing the restriction of a purely all pole model. This may be achieved by the use of parallel formant paths [1] or by the explicit introduction of zeroes in the transfer function. On the other hand, formant synthesisers generally use only three or four variable formants to characterise the speech spectrum. While this works well for vowel like sounds it can give rise to problems when representing non vowels. This is illustrated in Figure 2 and Figure 3 where the spectral response of a 7 pole LPC resynthesis is compared to a 4 formant response for the vowel [i:] and the fricative [v]. The formant parameters for these two sounds, as measured by the LPC root finding process described above, are also given in the figures. As can be seen from the figures the four formant synthesiser gives a good fit for the vowel but a poor fit for the fricative. This is primarily because the cascade synthesiser does not take into account all the poles in the transfer function, but only those with bandwidths below 500 Hz - ie the strongest poles. In the case of the fricative all the poles contribute approximately equally to the transfer function and so it is not possible to discard several of them even though they do fall outside our definition of a formant.

Generally speaking the formant synthesiser attempts to collect any non formant poles and zeroes together into spectral shaping filters associated with the excitation functions. For voiced sounds it is often stated that a low pass filter with a -12 dB/octave roll off can be used to represent the glottal spectral shaping and this should be combined with a high pass filter with a +6 dB/octave roll off, representing lip radiation. This simple approach may be extended [1,2] to include more complex voicing filter shapes. To investigate the appropriateness of fixed spectral shaping filters in a formant synthesiser we have examined a range of speech segments, located the formant positions with LPC root finding and computed the residual transfer function associated with the non formant poles. These spectral shaping functions are shown in Figure 4.

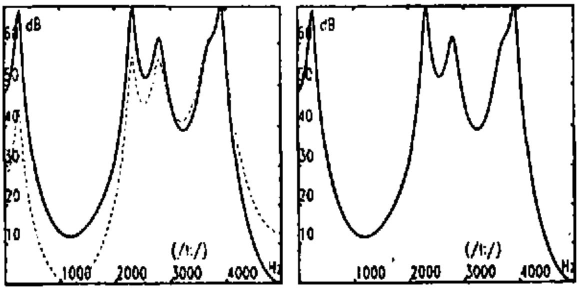
Formant poles

	F1	B1	F2	B2	F3	B3	F4	B4	F5	B5	(Hz)
[i]	265	84	2309	95	2806	183	3688	280	3932	111	
[æ]	648	312	1390	316	2534	269	3527	395	3774	190	
[ɑ]	525	112	963	156	2458	91	3638	85			
[ɛ]	599	92	1733	193	2603	155	3703	349			
[ɪ]	423	45	1694	177	2494	71	3620	322	3812	133	
[ɒ]	457	63	865	40	2390	226	3371	165			
[u]	366	48	1195	102	2243	129	3586	94			
[ɔ]	376	53	687	144	2465	134	3185	42			
[ʊ]	408	47	1149	34	2308	47	3609	142			
[ɜ]	497	73	1324	46	2420	65	3591	351	3764	119	
[ʌ]	624	81	1188	46	2403	75	3664	64			
[s]	3286	322									
[z]	347	76	1374	69	2378	193	3506	397	4314	446	
[θ]	1625	287	2447	349	4251	331					
[ð]	857	387	2273	381	2981	398	3652	241	4463	212	
[f]	1388	322	3060	345	3839	291					
[v]	144	65	3565	322							
[ʃ]	2393	275	3113	199	3767	188	4493	240			

Non formant poles

	F1	B1	F2	B2	F3	B3	F4	B4	F5	B5	(Hz)
[i]	0	768	1662	1310	5000	1765					
[æ]	0	396	1558	551	5000	352					
[ɑ]	1182	903	3200	804	4426	589					
[ɛ]	0	608	1533	806	3869	587	5000	412			
[ɪ]	0	1775	1551	890	5000	524					
[ɒ]	1558	767	3976	587	5000	2208	5000	503			
[u]	60	1918	3558	704	5000	1791	5000	405			
[ɔ]	1227	997	3693	589	4509	504					
[ʊ]	1227	961	3532	530	5000	1649	5000	427			
[ɜ]	0	933	1441	1520	5000	290					
[ʌ]	0	2185	0	835	2962	1123	4420	717			
[s]	0	223	1221	769	1573	812	2517	543			
	4226	508	4294	766	5000	364					
[z]	0	675	0	1233	4501	1493					
[θ]	0	60	655	1122	3229	523	3998	804	5000	3539	
[ð]	0	24	1430	1056	5000	1234					
[f]	0	227	0	679	2143	523	2431	192	4513	536	
[v]	0	4692	1199	545	2065	1024	2791	596	4419	683	
	5000	729									
[ʃ]	470	969	1378	991	2280	513					

Figure 1. Formant and non formant poles

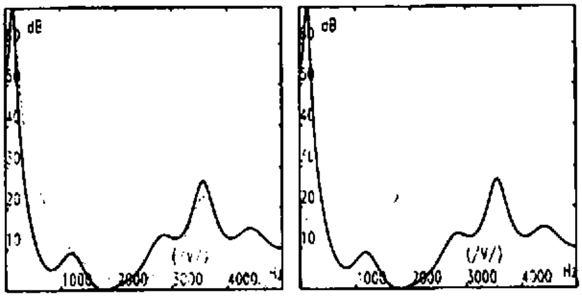


FORMANT PARAMETERS FOR THE VOWEL [I]

(Hz)	formants					non formants		
centre frequency	265	2309	2806	3688	3932	0	1662	5000
bandwidth	84	95	183	280	111	768	1310	1765

The left hand plot compares the all pole LPC spectrum (solid line) for the vowel [i] with the corresponding 4 cascade formant transfer function (dotted line). The right hand plot compares the LPC spectrum and the formant transfer function, when all the poles have been included in the formant transfer function. As expected, the two curves are coincident.

Figure 2. Comparision of 7 pole LPC resynthesis and 4 formant cascade synthesis

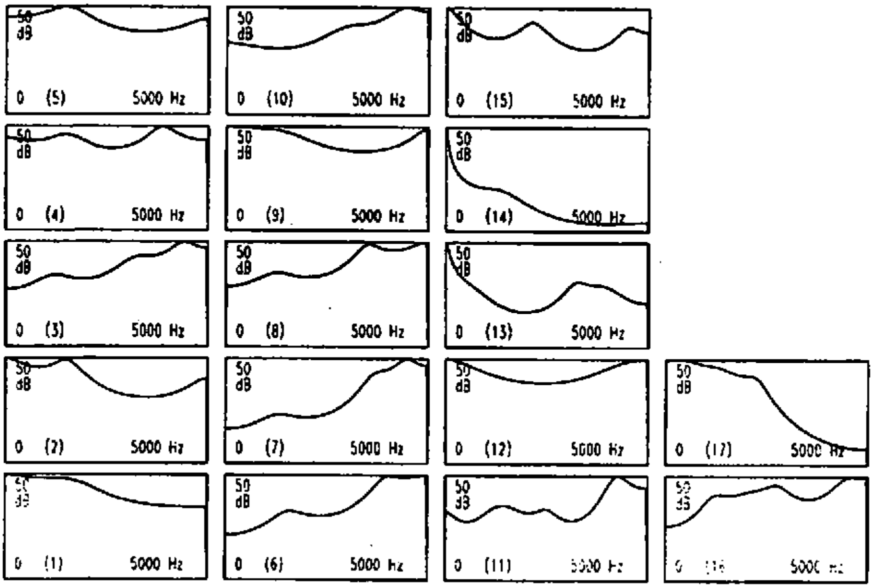


FORMANT PARAMETERS FOR THE FRICATIVE [v]

(Hz)	formants		non formants					
centre frequency	144	3565	0	1199	2065	2791	4419	5000
bandwidth	65	322	4692	545	1024	596	683	729

The left hand plot compares the all pole LPC spectrum (solid line) for the fricative [v] with the corresponding 4 cascade formant transfer function (dotted line). The right hand plot compares the LPC spectrum and the formant transfer function, when all the poles have been included in the formant transfer function. As expected, the two curves are coincident.

Figure 3. Comparision of 7 pole LPC resynthesis and 4 formant cascade synthesis



(1) [i], (2) [æ], (3) [ɑ], (4) [ɛ], (5) [ɪ], (6) [ɒ], (7) [ɔ], (8) [ʊ], (9) [ə],
(10) [ʌ], (11) [s], (12) [z], (13) [θ], (14) [ð], (15) [f], (16) [v], (17) [ʃ].

Figure 4. Spectral shaping due to non formant poles

It seems clear from these plots that the non formant spectral shaping is quite dependent on the particular speech segment being produced. It therefore seems reasonable that an improvement in the naturalness of a basic formant synthesiser could be achieved by accounting for this variability. This can be done by including the non formant spectral shaping in the voicing filters. In effect all the poles of the speech transfer function are then being taken into account by formant synthesis.

Adopting this strategy raises several questions:

1. If all the poles are accounted for in the formant synthesis process, is this not simply equivalent to performing an LPC resynthesis? This is in fact the case, but the retention of a formant synthesiser architecture has several advantages:
   a. zeroes may be readily introduced (for the synthesis of nasals for instance),
   b. the use of the familiar parameters of formant centre frequencies and bandwidths makes it easier to relate the synthesiser performance to the cosegmentation rules that will be used to drive it,
   c. classes of sounds may require the same spectral shaping, so that it may not be necessary to update the spectral shaping filters at as high a rate as it is necessary to update the formant resonators.
2. Is it not possible to account for any overall spectral shaping by adjusting the frequencies and bandwidths of the formant poles? This is indeed possible and can give

good results. However, this adds an extra level of complexity to the process of deducing synthesiser parameters from real speech.

On the other hand the introduction of variable voicing filters does have the disadvantage of increasing the complexity of the synthesiser, increasing the amount of computation required per frame of generated speech and of increasing the parameter data rate required by the synthesiser.

Resynthesis

If a piece of synthetic speech is to closely replicate an original human utterance it is clearly of crucial importance that there is a good spectral match between the two waveforms. However, it is also very important that sufficient attention is paid to the time domain aspects of the resynthesis. On the one hand care must be exercised in ensuring that the sampled formant parameter data is adequately smoothed over analysis frame boundaries otherwise spurious transients can be introduced into the synthetic speech, while on the other hand the rapid changes in formant parameters associated with stop closures must not be lost.

Conclusions

If synthetic speech is to sound acceptably natural it is necessary that the spectral qualities of the synthetic speech are well matched to those of actual human speech. The addition of time varying spectral shaping filters to the conventional formant synthesiser can provide addition controls that enable the spectral characteristics of the synthetic speech to be closely matched to those of natural speech. The control parameters for the spectral shaping filters are readily obtained from LPC pole location of natural speech. Retaining a formant based synthesiser architecture permits the further development of the synthesiser to accommodate zeroes in the speech transfer function.

References

[1] J.M. Rye and J.N. Holmes, 'A Versatile Software Parallel Formant Speech Synthesiser', Joint Speech Research Unit, Cheltenham, Report 1016, (1982).

[2] D.H. Klatt, 'Software for a parallel/cascade formant synthesiser'. JASA, 67 (3) p971-995, (1980).

[3] G. Fant, J. Mártony, U. Rengman and A. Risberg, 'OVE II Synthesis Strategy', Paper F5, Proc of the Speech Communication Seminar, Stockholm, Vol II, (1963).

[4] W.A. Ainsworth, 'Performance of a Speech Synthesis System'. Int J Man-Machine Studies, vol 6, p493-511, (1974).

[5] S.D. Isard and D.A. Millar, 'Diphone Synthesis Techniques'. IEE Conference on Speech Input/Output; Techniques and Applications, Conference Publication no 258, (1986).

[6] J.B. Pickering, 'Cosegmentation in the IBM Text-to-Speech System', IOA Autumn Conference, (1986).

[7] J.D. Markel and A.H. Gray, 'Linear Prediction of Speech', Springer Verlag, (1976).

[8] P.M. Seeviour, J.N. Holmes and M.W. Judd, 'Automatic Generation of Control Signals for a Parallel Formant Speech Synthesiser', IEEE Conference on Acoustics, Speech and Signal Processing, p690-693, (1976).

[9] L.R. Rabiner and R.W. Schafer, 'Digital Processing of Speech Signals', Prentice Hall, (1978).