Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 24 (2001), 167–187.

F_0 –RELATED FORMANT MEASUREMENTS

Ton G. Wempe

Abstract

The problems connected with the estimation of low formant frequency values of voiced speech sounds with high F_0 are widely known and stem from the fact that the distances between the frequency domain "samples" of the "filter function" of the vocal tract equal this high F_0. After some reflections about the limitations of the formant information in the signal itself and the influence of F_0, two different pitch–dependent analysis methods will be described. The first method (named 'Truncated Filtering Analysis') is pitch–synchronous and requires the F0 period to be isolated. The second (named 'Pitch–controlled Filter Analysis') can be applied to repetitive periods of the voiced speech and only needs the local pitch value. Spectra–producing scripts for the 'Praat' computer program that enable testing these two methods in practice are presented as well. Some measurements on artificial vowel–like sounds, performed in this way, are presented to check the accuracy in practice. Both methods produce spectra that form useful approximations of the spectral envelope which can be used as a basis for improved formant estimation, provided sufficient formant information is present in the signal itself. The latter method can be easily automated.

1 Introduction

Formant determination from the microphone signal of voiced speech sounds is still a difficult matter, despite its long history. The pitfalls are widely known: firstly, the voicing causes a sampled representation of the underlying spectral information so that high pitched speech sounds can give rise to a serious 'undersampling' of the spectral functions and secondly, the source function and the vocal tract filter function cannot be separated so that extracted formant data are based on some presumptions on both the source and the vocal tract properties. Naturally, there is no solution for the separation problem. We can only hope that the relatively fast closing of the vocal folds will produce evenly distributed and sufficient spectral energy to be able to map the spectral components into the vocal tract filter function. If not, we still have no choice but to regard the spectral data as representing the vocal tract filter function (after some pre–emphasis to correct for the expected source spectrum roll–off).

To tackle the undersampling problem many procedures have been developed, of which the LPC inverse filtering method is accepted as a reasonable compromise. However, the measuring parameters must be selected a priori, based on some properties of the speech itself. Some 'wrongly' chosen parameters could easily produce very misleading results (see Section 3 for some examples). Methods which can produce 'safer' results, therefore, can still be valuable and some will be presented below.

To gain some insight in the spectral properties of voiced speech and which information about the vocal tract function is available at all, the following sections deal with the basic properties of voiced speech and the consequences for the frequency domain so that the methods described next can be valued.

An old bandfilter–based analysis idea ('Truncated Filtering Analysis') was realized by us in the late 70's in the form of a hardware spectrum analyzer. The accuracy, analysis time and complicated operation of the equipment were the main drawbacks. The resulting graphs, however, were quite readable with respect to formant determination. At present, the computer program 'Praat' with its extreme flexibility can be used to simulate the old method and we made a script for testing purposes (see Appendix A). Section 4 describes its working principles.

Pitch–synchronous measurements cannot easily be automated, which limits the usefulness of the 'Truncated Filtering Analysis' design. Therefore an alternative sweeping bandfilter analysis ('Pitch–controlled Bandfilter Analysis') was tested which works on trains of periods instead. The choice of filter type and making the filter parameters dependent on the current pitch resulted in an analysis method that, at the cost of some frequency resolution, could be easily automated. Section 5 describes the method and Appendix B contains the 'Praat' script.

2 Formant information in voiced speech sounds

As a consequence of the periodicity of the (steady parts of) voiced signals, we have to deal with the problem of a sampled representation of a continuous envelope function (which comprises all formant information) in the frequency domain. Sounds with low formant–to–fundamental ratios (female and infant voices), therefore, hinder the estimation of these formants and the separation of formants close to each other. The periodic property of the signal means that all available information about the form of this envelope is contained in each of the fundamental periods of the steady part. (Scanning the envelope by using possibly varying F_0 to get more envelope information is no real solution in practice, because there is insufficient probability that the articulation movements don't cause too much formant changes within the sweep time. For the same reason we cannot rely on noise excitation because in that case we need many different parts of the speech for sufficient averaging.)

In source–filter terms we may consider the waveform of one F_0 period to consist of a sum of damped sinusoids, all starting anew again in the next period. The final amplitude of a damped sinusoid at the end of the pitch period equals $A \exp(-\alpha T_0)$, where A is the initial amplitude and T_0 is the length of one period of F_0. For the moment adopting the simplification to consider the damping α as being a fixed part of the formant frequency (constant Q factor of the vocal tract 'bandfilters'), enables us to regard the final amplitude as being proportional to the formant–to–fundamental ratio (F_F/F_0).

Although for high formant–to–fundamental ratios (mostly male voices) we may neglect the filter energy at the end of the F_0 period, this is not necessarily the case at lower ratios (i.e. female and infant's voices). The filtering process can be seen as the multiplication of the source spectrum (discrete spectrum with multiples of F_0, gradually decreasing in amplitude with increasing frequency) and the filter 'spectrum', so that, after some pre–emphasizing to correct for the spectral slope of the source, the speech signal is nothing else than the filter function sampled in the frequency domain at F_0 distances, *independent* of the amount of energy at the end of the F_0 period. This filter function, therefore, is the spectrum of one F_0 'period' if it were left undisturbed till infinity, in other words the vocal tract filter impulse response. The way the filter function area is ’filled’ with samples, however, is dependent on the positions of the samples in relation to the filter curve. (See Figure 1 for an example with artificial vowel–like sounds with low formant–to–fundamental ratios and their spectra.) Extremes are formed in cases where the formant frequency equals a multiple of F_0 (Fig. 1A) and where the frequency falls midway between two adjacent spectral lines (Fig. 1B).

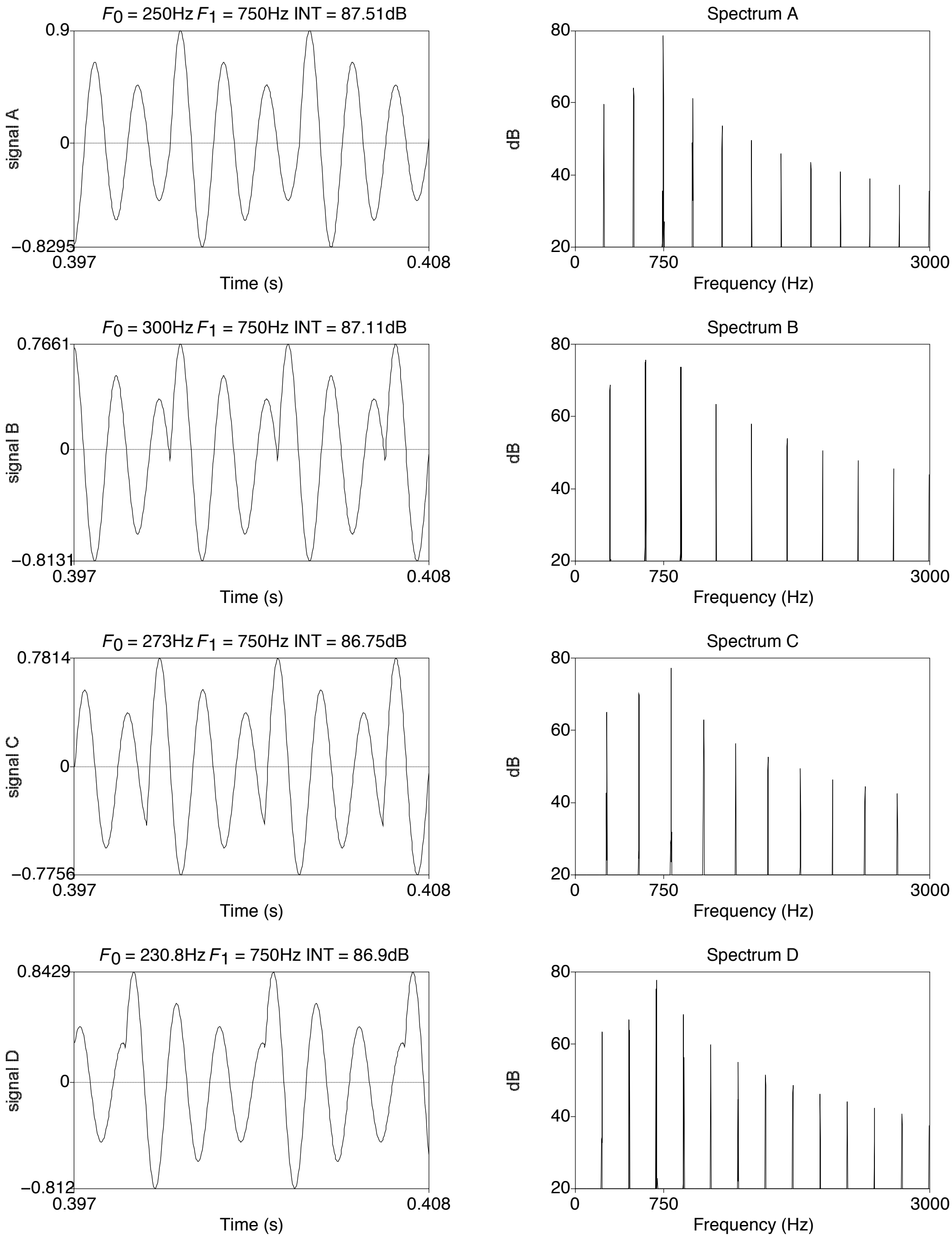

Figure 1. Four different ’one formant’ signals and their long–term spectra. The formant frequency for all signals is 750 Hz. F_0 varies so that the number of damped sinusoid periods within one fundamental period in signal A is 3, in signal B is 2.5, in signal C is 2.75 and in signal D is 3.25. No formant frequency shifts and almost no initial phase changes can be seen. The maximum intensity difference is only 0.76 dB.

The consequence is that the amplitude of the damped sinusoid is somewhat dependent on F_0 (the frequency and initial phase of the damped sinusoid remain practically unaltered). The intensity differences caused by this effect, however, are roughly below 0.75 dB (*in the stationary part*) for all practical vowel sounds. Generally this effect, therefore, can be neglected in practice.

Regarding the spectral envelope as the filter function of the vocal tract implies that the source signal is to be considered as a stream of delta pulses (or at least pulses shorter than, say, 0.1 ms so that the first zero of its $\sin(\pi f T)/\pi f T$ spectrum lies as far away as 10 kHz). Of course in reality the pulse is shaped differently. However, our only information source for formant estimation is this spectral envelope. We assume then that the relatively fast closing of the vocal folds form harmonics that gradually decrease with increasing frequency, without local maxima or minima. Pre–emphasizing can correct for the roll–off to some extent.

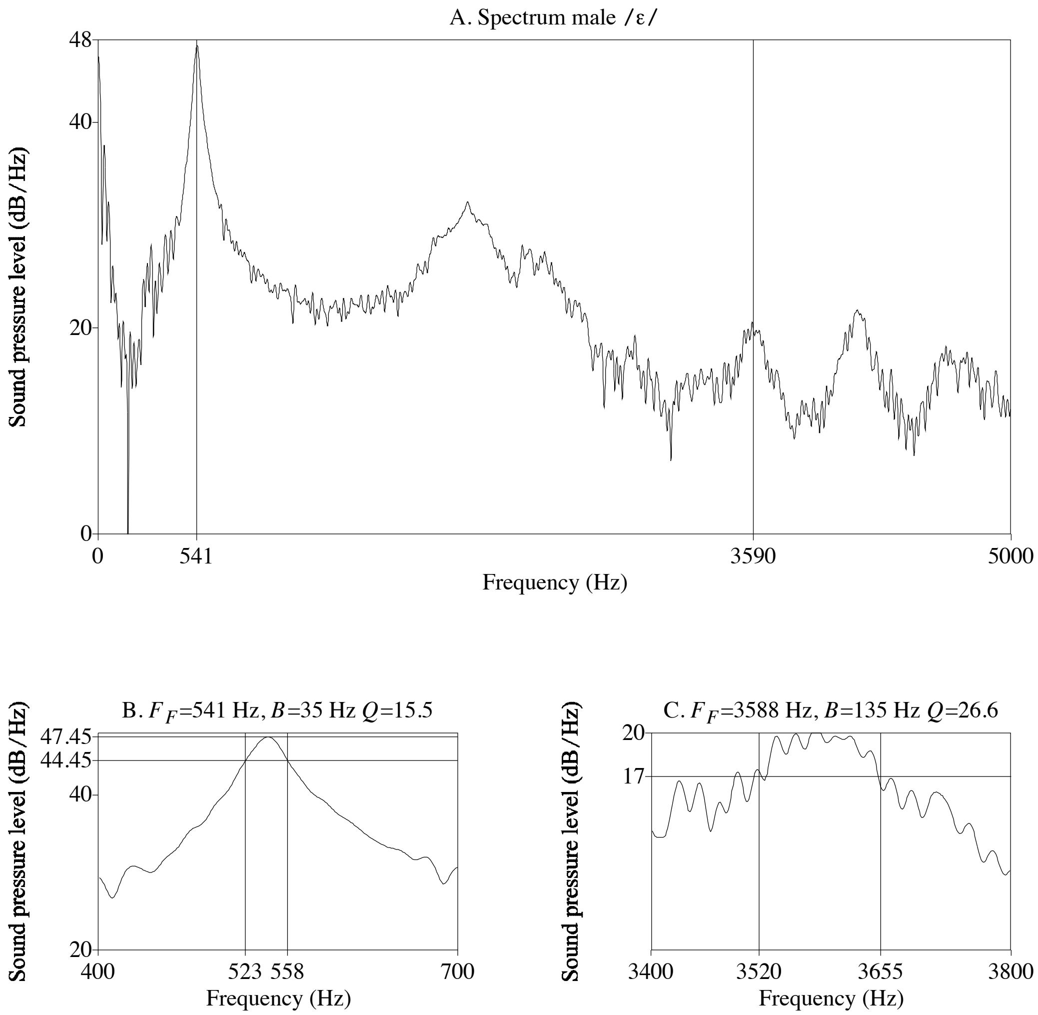

Figure 2. Low–F_0 spectrum of a male /ɛ/ (53 ms period of a creaky voice, with appended 0.45 s zero sound) which shows rather high Q values. Probably the Q–factor of the low formant (B) is even higher in the real envelope spectrum because of the 53 ms window spectrum convolution.

The sampling theorem limits the maximum frequency that can be regained from a sampled time function to half the sample frequency. The same applies to the frequency domain which roughly means that it is impossible to distinguish from the spectral envelope two peaks less than $2F_0$ Hz apart and that the determination of the position of a peak implies an inaccuracy of more or less F_0 Hz. Of course that applies only to a completely unknown form of the envelope. Generally it is assumed that the envelope changes are quite smoothly because the Q–factors (formant frequencies devided by their bandwidths) of the resonances of the vocal tract are considered to be rather limited, and the number of peaks to be low as well, so that, in spite of the undersampling, the envelope can be approximated reasonably well. However, depending on the open/close ratio of the vocal folds movements, the Q–factor can vary a great deal. Some pilot measurements (on low F_0 voices) learn that Q can easily reach values like 20 or 25 (see Figure 2 for an example with a 19 Hz F_0).

Therefore, when the filter function is undersampled the reconstitution can never approximate the sharper peaks of the original function very accurately: they become broadened. The bandwidths of the peaks of reconstituted envelopes in that case are heavily dependent on F_0. A formant peak of say, 540 Hz, having a bandwidth of less than 27 Hz, can only be approximated satisfactorily when F_0 is a great deal lower than about 27 Hz. In practice that would be highly exceptional (hence the creaky voice example). Bearing in mind that we don't know the number of prominent formant peaks either, the undersampling effects will seriously limit the accuracy of formant extraction, regardless of the type of analysis: the adequate information then is simply not present in the signal.

3 Formant extraction

Obviously, an attempt to approximate the envelope by calculating the (continuous) spectrum of one isolated F_0 period offers no solution for low formant–to–fundamental ratio signals: the truncated period can be regarded as a multiplication of the untruncated time signal (i.e. the impulse response of the vocal tract filter) and a rectangular time window with length T_0. The result in the frequency domain is the convolution of the filter spectrum with the $\sin(\pi f T_0)/\pi f T_0$ spectrum of the time window so that the side lobes of the window function show minima and maxima at F_0 intervals. The (under)sampling problem of the envelope is back again!

Alternative window functions could be applied to decrease the 'side lobe ripple' amplitude. However, the attenuation of the signal parts outside one period must be sufficient which means that a substantial part of the signal within the period will be attenuated as well. The side lobe suppression is then realized at the cost of frequency resolution. Furthermore, the position of the window center must coincide with the center of the period (pitch synchronous windowing).

In order to estimate formant frequencies many strategies have been developed. Apparently, from all these methods the LPC inverse filtering seems to be considered as the de facto method for extracting formants nowadays.

Although LPC has its advantages, these are mainly in terms of the (often welcomed) reduction of the spectral envelope complexity and the direct numerical presentation of peak data, instead of reliable approximation of the filter function of the speech sound. The main drawback is the requirement to define the order of the LPC analysis in advance, based on several presumptions about the signal properties. When the order selection is made inappropiately, spectral peaks can easily emerge in entirely the wrong places (see Figure 3 for some unfortunate LPC spectra of artificial vowel–like sounds). The rules–of–thumb for selecting the number of peaks and frequency range limits dependent on male or female voices are not always appropriate, especially when the signal parameters are not well–known (infant's voices?).

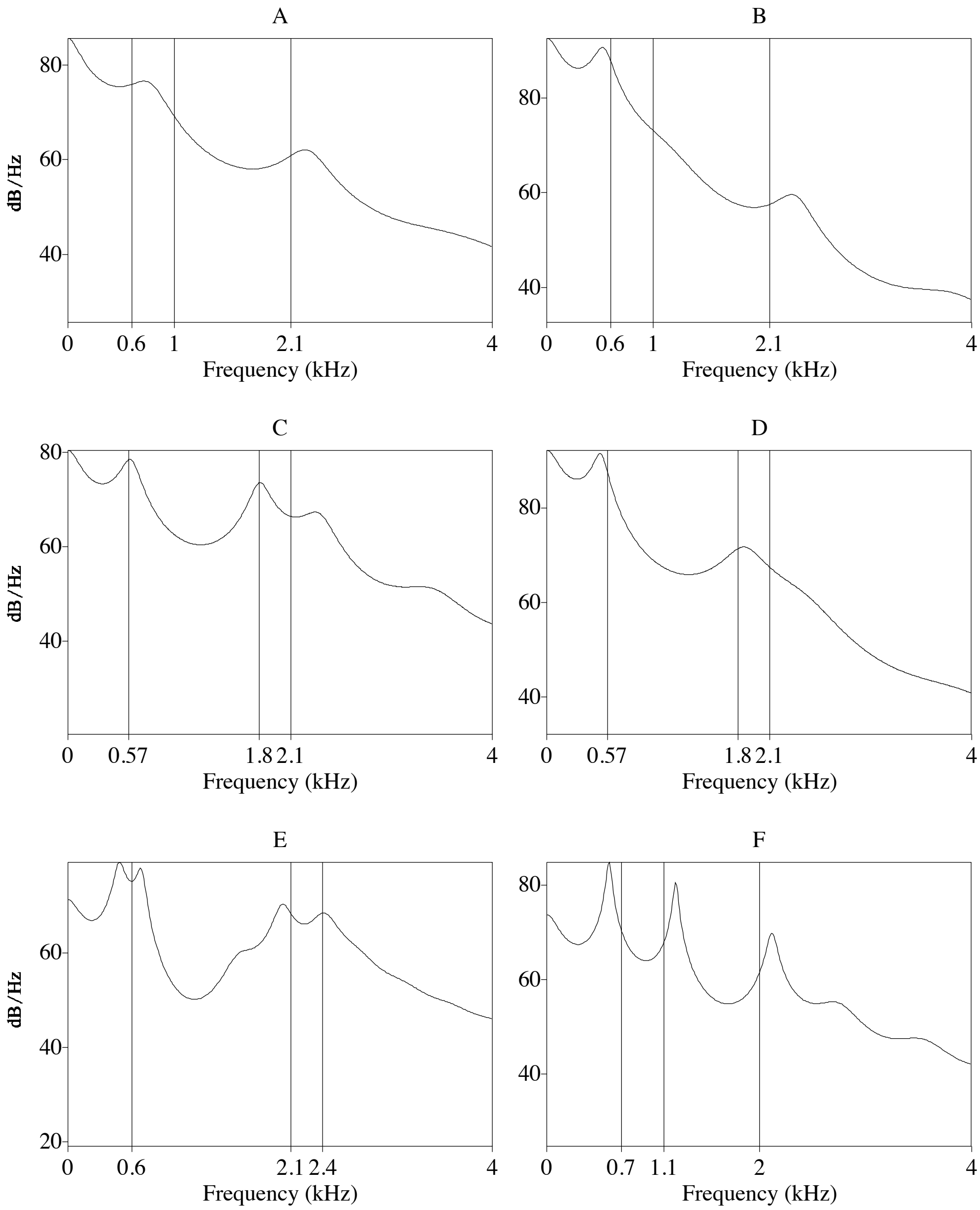

Figure 3. Some LPC spectra of various artificial vowel sounds made with 'unfortunate' but not 'insane' parameter selections (see Table 1). The dotted lines mark the generated formants. For all formants: Q=10. A de–emphasis of 6 dB/octave from 3kHz on was applied to all signals for elimination of possibly high frequency energy caused by amplitude steps of the period crossings of the artificial sounds used (see note on page 179 for a remark on the signals used).

Table 1. The parameters of the artificial vowel sounds and the LPC spectra from Fig. 3. All sounds were de–emphasized from 3 kHz to simulate a normal roll–off and resampled to 10 kHz, prior to the LPC analysis.

Sig.	F_0	A_1	F_1	A_2	F_2	A_3	F_3	Pre–emp.	LPC order
A	180	0.5	600	0.3	1000	0.2	2100	y	8
B	180	0.5	600	0.3	1000	0.2	2100	n	8
C	220	0.4	570	0.3	1800	0.3	2100	y	10
D	220	0.4	570	0.3	1800	0.3	2100	n	10
E	170	0.4	600	0.3	2100	0.3	2400	y	24
F	300	0.4	700	0.4	1100	0.2	2000	y	12

In addition, some vowel sounds have substantially more prominent peaks than others within these ranges, and we may also be interested in higher frequency areas. Another disadvantage is the lack of information about bandwidths in the LPC spectrum. Some peaks can be broad and some can be razor sharp. Furthermore the LPC's susceptibility to the spectral slope differences of the signals can present problems because the quality of speech recordings can vary a great deal in this respect. Finally the LPC spectrum happens to be quite noise dependent. There exists some suspicion therefore that in practice most researchers make spectrograms prior to running LPC analyses in order to see where and how prominent the formants are and which LPC parameters to select!

4 Pitch–synchronous methods

4.1 The origin of the side lobes

Firstly, to gain some insight in the spectral properties of damped sinusoids with finite duration, we will adopt a somewhat unusual approach and consider a truncated damped sinusoid as the result of the subtraction of two untruncated damped sinusoids:

$$g(t) = g_0(t) - g_m(t) \tag{1}$$

where $$g_0(t) = \exp(-\alpha t)\sin(\omega_F t) \qquad (0 < t < \infty) \tag{2}$$

and $$g_m(t) = \exp(-\alpha t)\sin(\omega_F t) \qquad (T_0 < t < \infty) \tag{3}$$

Here is ω_F the frequency of the (damped) sine wave. See Figure 4 for the form of $g_0(t)$ and $g_m(t)$.

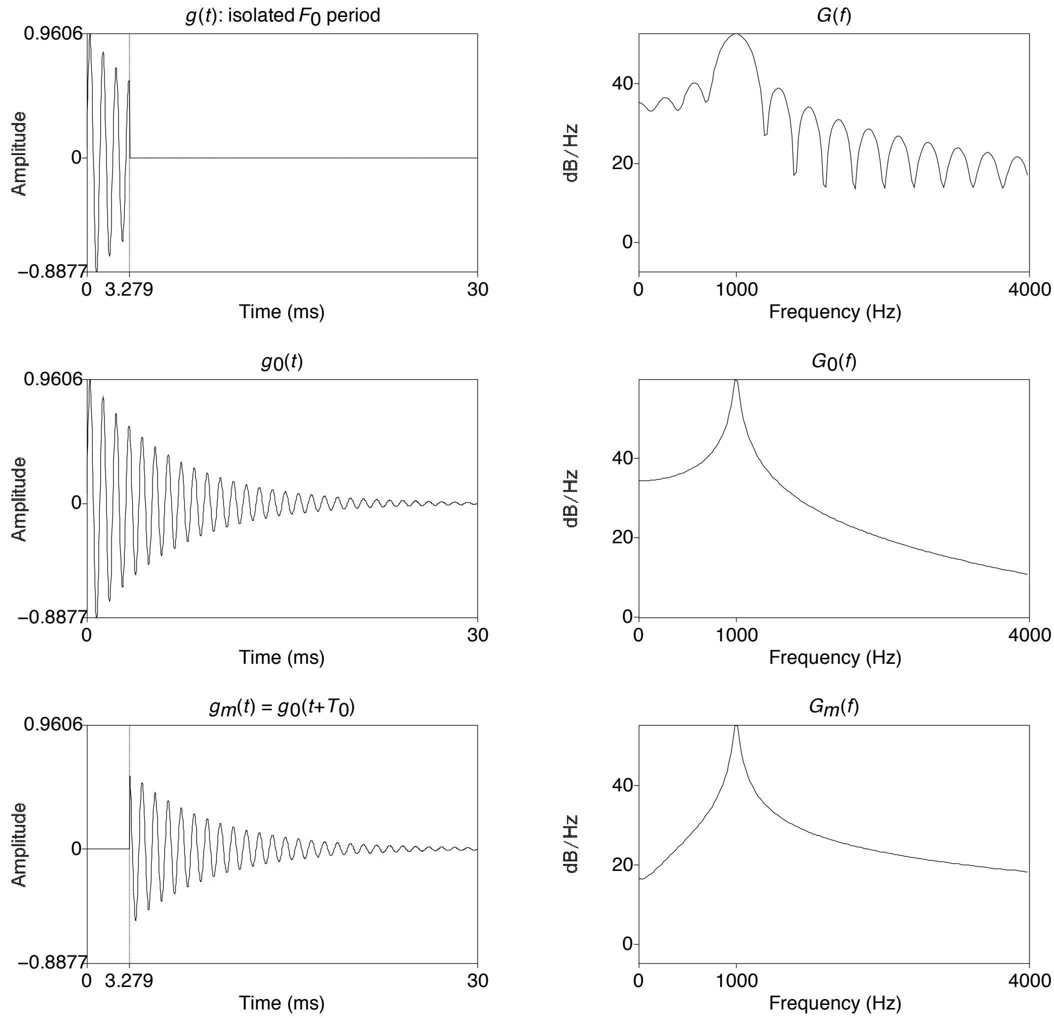

Figure 4. The truncated damped sinusoid of a 'one formant' F_0 period can be seen as the subtraction of two untruncated damped sinusoids: $g(t) = g_0(t) - g_m(t)$. In the frequency domain this is equivalent with the untruncated 'original' spectral components minus their corresponding time shifted spectral components, while accounting for the relative phase differences, which increase linearly with the frequency.

The (modifying) function $g_m(t)$ can be regarded as a phase shifted version of the impulse response of the (one−formant) vocal tract filter, attenuated by the factor $\exp(\alpha T_0)$, and time−shifted by T_0. The (continuous) spectrum of $g_0(t)$ is:

$$G_0(\omega) = \frac{\omega_F}{(\alpha + j\omega)^2 + \omega_F^2} \tag{4}$$

and its (squared) amplitude spectrum:

$$|G_0(\omega)|^2 = \frac{\omega_F^2}{(\alpha^2 + \omega_F^2 - \omega^2)^2 + 4\alpha^2\omega^2} \tag{5}$$

If $\omega_F^2 >> \alpha^2$ (which generally is the case) the function $|G_0(\omega)|$ exposes a maximum at ω_F and a −3dB bandwidth of α/π Hz.

The spectrum of $g_m(t)$ can be written as:

$$G_m(\omega) = \exp(-j\omega T_0)\cdot\exp(-\alpha T_0)\cdot G_0(\omega) \tag{6}$$

The factor exp(−?T_0) defines its initial amplitude and the function exp(−*j*?T_0) expresses the phase differences with the G_0 components which increase linearly with increasing frequency, according to the time−shift property of the Fourier transform. If the *initial* phase of the modifying function is *n?* (damped sine), the amplitude spectrum of $g_m(t)$ (not its complex spectrum because of the increasing phase differences) simply is an attenuated version of the amplitude spectrum of $g_0(t)$:

$$|G_{m1}(\omega)| = \exp(-\alpha T_0)\cdot|G_0(\omega)| \tag{7}$$

If the initial phase equals $\pi(1/2+n)$ (damped cosine) the amplitude spectrum becomes:

$$|G_{m2}(\omega)| = \exp(-\alpha T_0)\cdot\left|\frac{\alpha + j\omega}{(\alpha + j\omega)^2 + \omega_F^2}\right| \tag{8}$$

which can be written as:

$$|G_{m2}(\omega)| = \exp(-\alpha T_0)\cdot\sqrt{\frac{\alpha^2 + \omega^2}{\omega_F^2}}\,|G_0(\omega)| \tag{9}$$

The factor $\sqrt{(\alpha^2+\omega^2)/\omega_F^2}$ causes a 6 dB per octave 'pre−emphasis' with respect to $G_0(\omega)$. (See Figure 5A for the amplitude spectra differences.)

The superposition property of the Fourier transform means that the subtraction of the time functions can be performed by subtracting all separate corresponding frequency components, while accounting for the relative phase differences. Because of the fact that the phase differences increase proportional with the frequency, maximum deviations from the spectrum of $g_0(t)$ occur when corresponding components have equal or opposite phases. (In case the amplitudes of both functions are the same, zeroes occur at F_0 distances from the 'formant' peak, and +6dB deviations in between, which is in agreement with the $\sin(\pi fT)/\pi fT$ spectrum of a truncated sinusoid.) The lower the modifying function amplitude, the weaker the 'ripple'. However, it is still possible that the truncated damped sinusoid spectrum exposes (near) zeroes at a specific local frequency area, as explained below.

Depending on the initial phase of the modifying function, its spectral slopes can differ from those of $G_0(\omega)$ (see Figure 5A). Only when the initial phase of the modifying function equals that of the primary function, both spectra have the same form. In that case the 'ripple amplitude' as a function of frequency follows the amplitude spectrum of $g_m(t)$. When the spectral slope of the modifying function is less steep than that of the primary function, they intersect at some frequency and, at equal phases, may cancel each other. See Figure 5B where the primary function has zero initial phase and the modifying function $\pi/2$ radians.

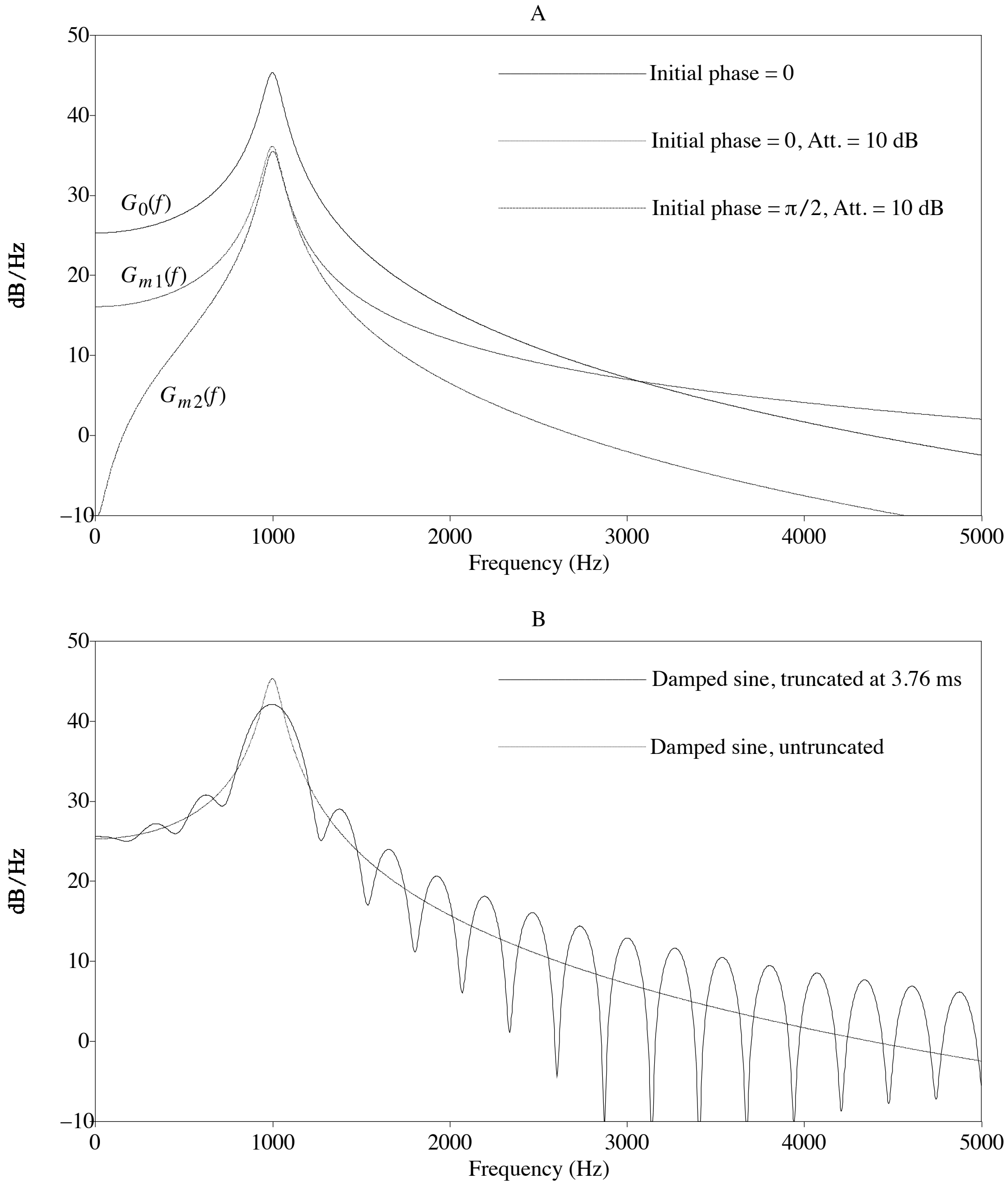

Figure 5. The spectral slopes of damped sinusoids are dependent on the initial phase. When the spectral components of the 'modifying' function G_m are subtracted from those of the 'primary' function G_0, the different slopes will cause a cancelling area where the weaker modifying function intersects the stronger primary function (A). In this example the resulting spectrum of the truncated damped sinusoid shows (near) zeroes in the vicinity of 3000 Hz (B).

To conclude, we find that in the spectrum of one F_0 period the influence of F_0 on the magnitude of the 'side lobe ripple', the forms of leading and trailing slopes of formant peaks, and their bandwidths, is substantial and should not be ignored. In addition, we see that *only the modifying function* $g_m(t)$ is responsible for the side lobe occurrences.

One way to decrease the ripple in the spectrum is the multiplication of the F_0 period with an exponential window $g_w(t) = \exp(-\beta\ t)$ where β is chosen such that the final amplitude of the period T_0 is negligible (i.e. the amplitude of the modifying function). All side lobe peculiarities will then vanish in practice (except that the spectral slope still depends on the initial phase). Of course the spectral peaks are broadened because of the convolution of the window spectrum $G_w(\omega) = 1/(\beta+j\omega)$, which has a bandwidth of about β/π Hz, with the signal spectrum $G(\omega)$. Although the readability of formants can improve considerably (see Figure 6), the drawbacks are the necessity to work pitch synchronously and to adjust β dependent on F_0, formant frequency and damping.

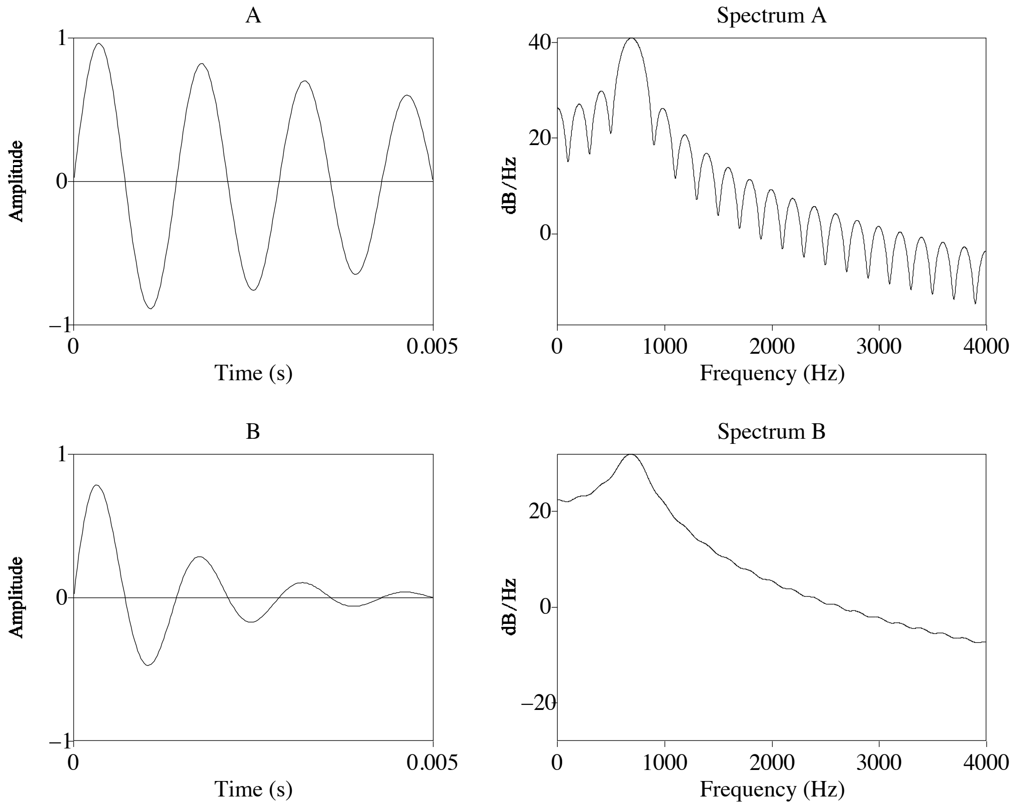

Figure 6. A: One period (5 ms) of a 700 Hz formant signal and the spectrum of the isolated period. B: The same after multiplication with an exponential window exp(−600t). The side lobes are strongly reduced.

4.2 A special pitch−synchronous method: the 'Truncated Filtering Analysis'

Bearing in mind that the F_0 effects in the spectrum are caused entirely by the influence of the modifying function $g_m(t)$, we can think of ways to minimize that influence. When we look at the time domain output of an analyzing bandfilter, for instance, we see that the response of the filter during T_0 is built up from zero value. At the end of the period the filter output is "disturbed" by the next period (or by the abrupt ceasing of the signal when the period was isolated). Now, when we measure the energy of the filter output *only during the* T_0 *interval*, we omit the influence of the modifying function completely and get a value which is a function of the bandfiltered spectral energy of the *untruncated* vocal tract impulse response. The frequency range of interest can now be scanned in small steps, making sure that the filter always starts from zero energy. The bandwidth can be chosen to be much smaller than F_0 (for there is no need to bother about the periodicity) which means that the frequency resolution is mainly restricted by the signal itself and not by the analyzing method.

Of course in that case the filter transmission time is greater than T_0 so that the output at the end is far from the steady state. The envelope form of the filter output built–up, however, is only dependent on the filter type and bandwidth (which remain unaltered during the analysis) and is scaled by the filter value at the current frequency. Therefore the output energy values can be regarded as being proportional with the spectral energy of the filtered part of the sound. Figure 7 shows some consecutive filter output steps in the vicinity of the (single) spectral peak of a signal (1000 Hz damped sine).

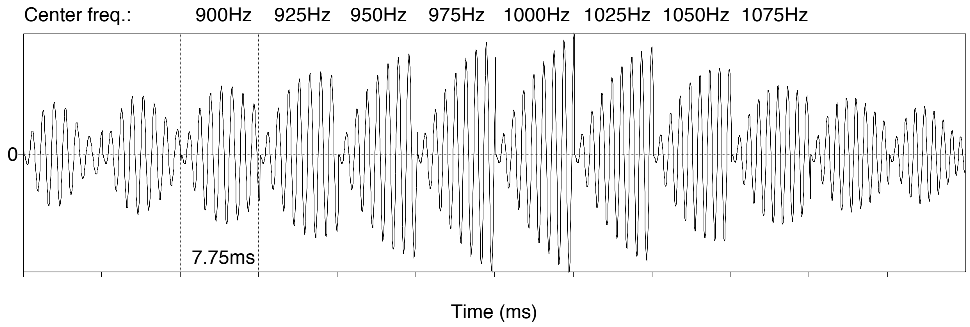

Figure 7. Train of consecutive 'Truncated Filtering Analysis' filter outputs in the vicinity of the spectral peak in the signal (1000 Hz damped sine truncated at 7.75 ms).

Some restrictions exist: firstly, at very low frequencies the peak heights can deviate slightly because the energy of a low number of sine halves within the T_0 interval is somewhat dependent on that number and secondly, the bandfilter order should be low to prevent complicated time behaviour of the filter impulse response.

In practice the first restriction is negligible: the power of a truncated sine wave is

$$P = \frac{1}{T_0}\int_0^{T_0} \sin^2(\omega t)dt = \frac{1}{T_0}\int_0^{T_0} \frac{1}{2} + \frac{1}{2}\cos(2\omega t)dt = \frac{1}{2}\left(1 - \frac{\sin(2\omega T_0)}{2\omega T_0}\right) \tag{10}$$

(which is the sine integral of the double frequency devided by T_0) so that maximum deviations occur at $(n+1/2)/4$ periods of the sine wave within T_0. When, for example, $n = 6$, i.e. the sine frequency is only 1.625 times the 'fundamental' ($1/T_0$), the ripple is less than 0.5 dB. Of course this effect is even lower for *damped* sinusoids.

The second restriction stems from the fact that the time output of the filter is formed by the convolution of the impulse response of the current filter (which is a damped sinusoid at its current center frequency) and the signal period. Generally, the higher the filter order, the more complicated its impulse response. In addition, using a filter of which the impuls response is a damped sinusoid is to be preferred, which is explained below.

Although the function of two convolved signals differs from their cross–correlation function, the *amplitude* spectra of the convolution and of the correlation are equal, so that this analysis method can be thought of being a set of cross correlations of the signal and the filter impulse responses at each frequency step. In principle, when estimating formants, we want to look for peak correlation values of the signal F_0 period and *truncated* damped sinusoids, whereby the frequency which causes the highest correlation corresponds with the formant frequency. Therefore it makes sense to select a filter which has a damped sinusoid as its impulse response, i.e. a 2nd order bandpass filter. (When we time–reverse one of the time functions, the filter output is exactly equal to the cross correlation.)

The (gradual) building up of the filter outputs during the T_0 intervals can thus be made identical for all center frequencies by applying a constant bandwidth filtering (which means optimal resolution at higher formant frequencies as well) so that the obtained intensity values are proportional to the real spectral values. Furthermore, making the bandwidth selection proportional to F_0 enables comparison of spectral graphs from signals with different F_0 values.

A hardware spectrum analyzer based on this 'Truncated Filtering Analysis' principle was made as early as 1979 (Wempe, 1979). A "Praat" script which simulates this hardware analyzer is presented in Appendix A, together with a global explanation.

Figure 8 shows the resulting power spectra for some artificial vowel sounds[1] so that the accuracy and prominence of the spectral peaks can be judged. The tendency to shift low frequency peaks to the lower end of the frequency axis, which is a 'natural' property of a damped sine spectrum, could easily be corrected automatically, because the correction factor for each frequency can be derived from the current central frequency of the bandfilter.

[1] Note on artificial signals used. Naturally, the discontinuities at the end of the F_0 period of the applied artificial signals don't occur in reality. According to the cascading filter concept of speech sounds, the F_0 period boundaries occur rather smoothly. To test the spectral analyzing methods, however, it is quite convenient to be able to define the signal parameters of the test signals independently of each other, which is not possible in the case of the cascading filter concept. Particularly for the spectral peaks there is no great fundamental difference: a gradual roll–off at high frequencies could simulate the cascading rather well. Therefore a 6 dB per octave de–emphasis from 3000 Hz on was applied to all signals.

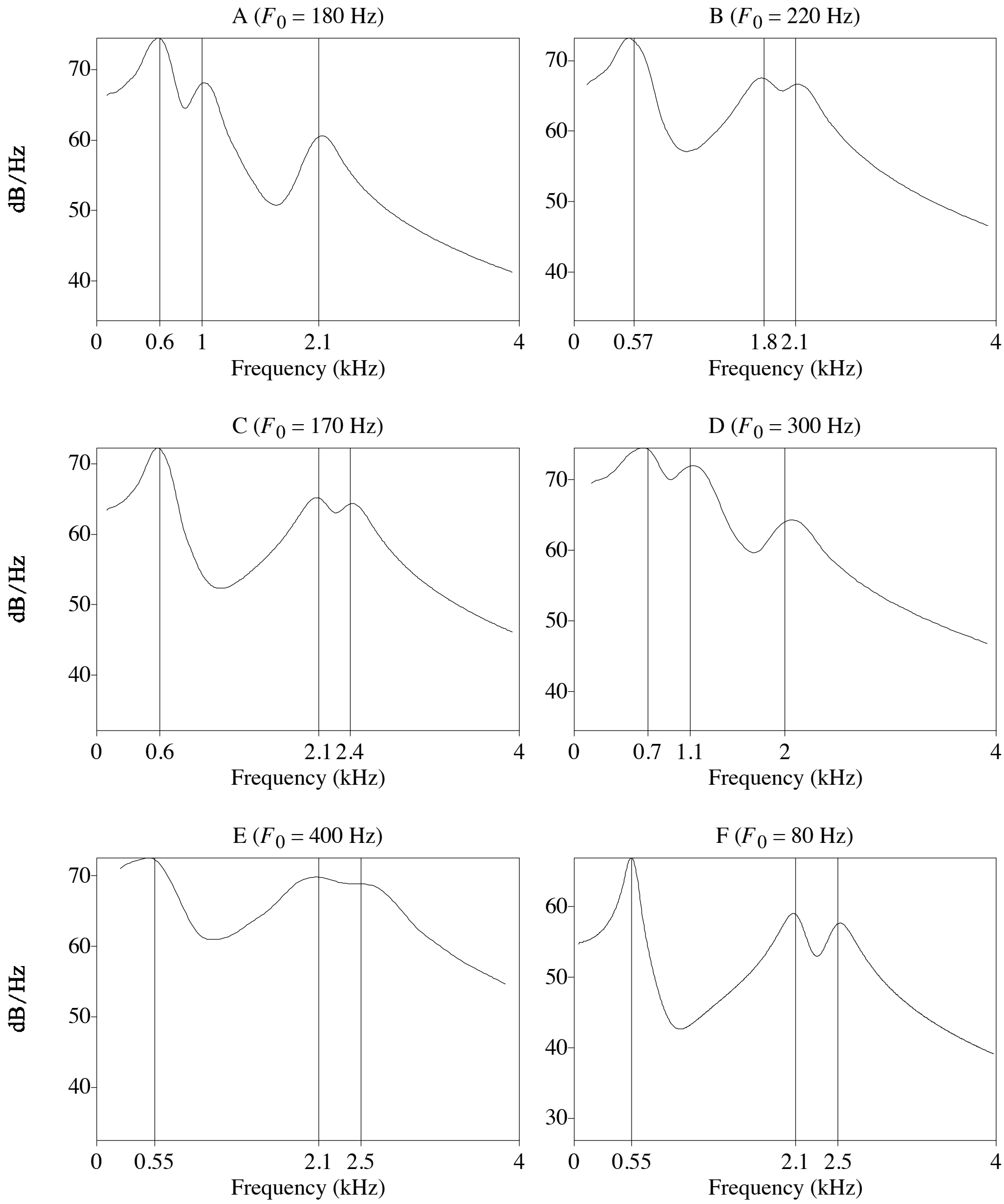

Figure 8. Power spectra obtained with the 'Truncated Filtering Analysis'. Graphs A through D show spectra from the same artificial signals as used for the LPC spectra in Figure 3, except that they are all sampled at 44.1 kHz. Graph E shows that a very low formant with respect to F_0 can be detected, as well as two formants that differ only F_0 Hz. Graph F shows that in practice the F_0 value has an influence on the peak widths only.

5 The 'Pitch–controlled Bandpass filter Analysis'

Although the properties of the previous method are quite attractive, the main drawback is the necessity to isolate the F_0 period of the speech signal and to find the position of the origin of the damped sine waves. Especially in cases where the first formant to F_0 ratio (F_1/F_0) is low (female and infant's voices) these period boundaries are difficult to find. The achieved accuracy can easily become insufficient when automation of the process is attempted: the spectrum of a wrongly isolated period (thus containing a 'phase step') has not much to do with the vocal tract filter function.

If we wish to avoid the determination of the F_0 period crossings we have to deal with *trains* of F_0 periods. The impulse response of any filtering method then should have a suitable form for windowing, i.e. for minimizing the 'interperiodic' influence. When analyzing a periodic vowel signal with a swept bandpass filter, the time domain output in each frequency step is formed by the convolution of the vowel signal with the bandfilter impulse response.

An ideal (rectangular) filter with bandwidth B has an impulse response of the $\sin(x)/x$ form where the zeroes are positioned $1/B$ seconds apart. When the filter bandwidth is equal to F_0, the side lobes of the impulse response coincide exactly with the repetitive periods of the vowel sound and all give a weighted contribution to the convolution values. The final result is that the spectral envelope has been approximated with a staircase function where the step widths are F_0 Hz (as can be expected from a properly reconstituted sampled function). Decreasing the filter bandwidth introduces zeroes again and increasing the bandwidth will deteriorate the frequency resolution. It will be clear that no improvement has been achieved with respect to the discrete frequency samples of the Fourier spectrum.

Appropriate forms of the impulse response of the measuring filter should be such that its energy after T_0 seconds has decreased sufficiently if the influence of the next signal period is to be minimized: the convolution with the individual signal periods should not interact too much. A low–order bandfilter with sufficient bandwidth, identical to the classical broadband filter analysis, could perform the task. After all, the amplitude of the impulse response (damped sinusoid) at the end of the F_0 period can be controlled by the choice of bandwidth.

A second–order bandpass filter has the advantage that its filter function (which is similar to the spectrum of a damped sine wave) has one prominent peak and its gradual attenuation at both sides from its center frequency means that many spectral components from the vowel spectrum play a part in the response to the signal. The spectral peaks can be presented with relatively high resolution whereas the valleys are smoothed. These properties make it possible to suppress the ripple substantially.

Assuming constant *percentage* bandwidths (constant Q factor) of the formant peaks of the envelope function, the final amplitude of a high formant is much lower that that of a low formant (final amplitude $A_E = \exp(-\alpha T_0)$ where α is proportional to the formant frequency). It seems, therefore, that the measuring filter bandwidth could be decreased with increasing central frequency to gain frequency resolution for higher formants. However, when two formant frequencies are Δf apart, they can only be distinguished when the T_0 interval is greater than $1/\Delta f$ (the available time interval must be sufficient to contain the low 'period' of Δf). Obviously, the optimal bandwidth has to be proportional to F_0.

For 'difficult' signals where the formant frequency falls midway between two spectral lines, B has to be 1.5 F_0 or greater for a reasonable suppression of the spectral side lobes. In practice the choice $B = 1.25\ F_0$ turns out to be a proper overall compromise.

Making the analysis dependent on the current local pitch can be realized quite easily: the pitch detection in 'Praat' offers very reliable data and, besides, there is no need to localize the period crossings.

A "Praat" script which analyses an isolated period in this way is presented in Appendix B with a global description. Figure 9 shows some spectra obtained with this method for some artificial vowel sounds. The applied (complex) filter function has the form:

$$G_{BF}(f) = \frac{jfB}{f_R^2 - f^2 + jfB} \tag{11}$$

where f_R is its center frequency and B its −3 dB bandwidth. This function is preferred as it is symmetrical on a log frequency scale, unlike the spectra of damped pure sine or cosine waves. Its impulse response is a damped sine wave as well, the initial phase, however, is slightly less than $\pi/2$ and somewhat dependent on the bandwidth.

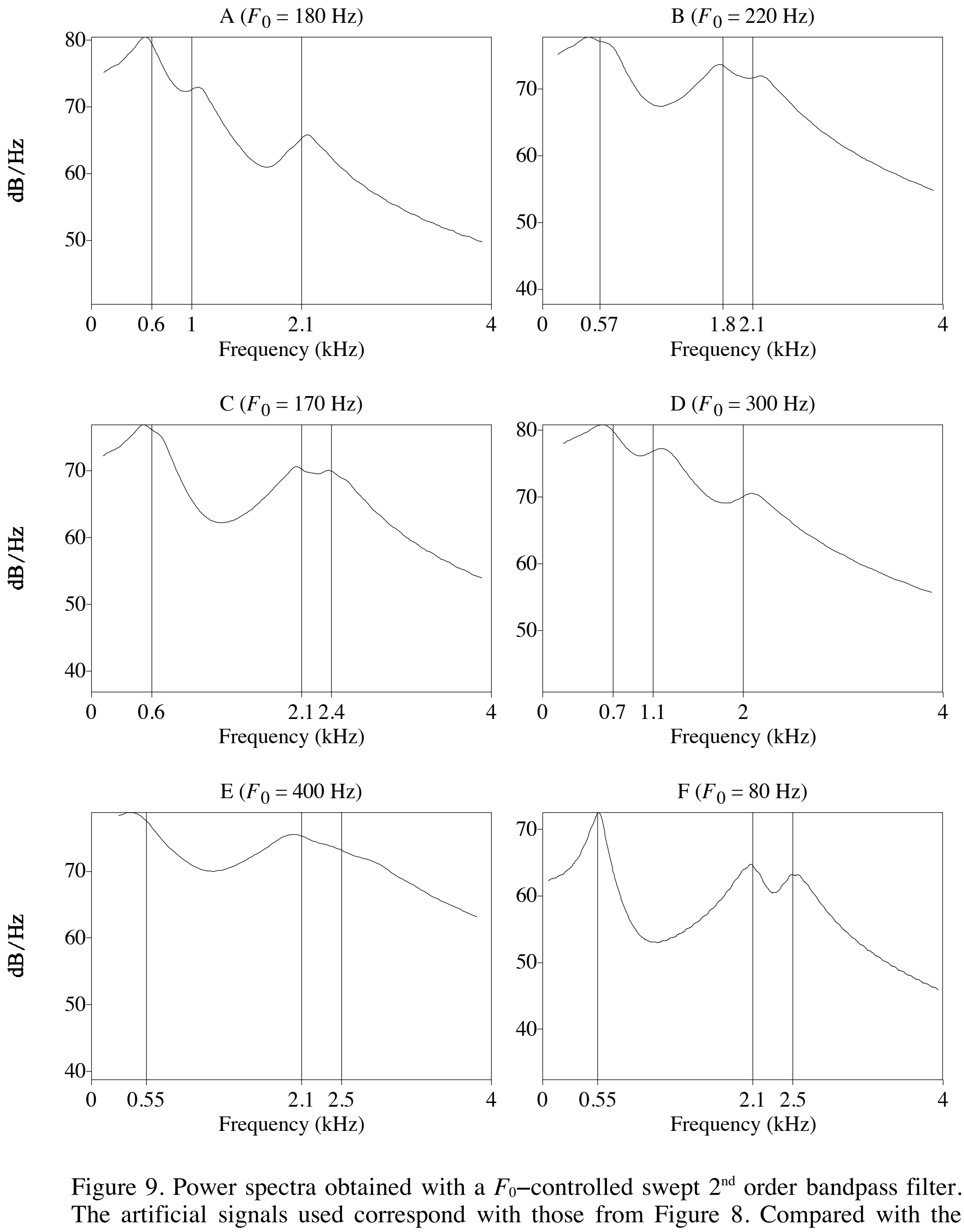

Figure 9. Power spectra obtained with a F_0–controlled swept 2nd order bandpass filter. The artificial signals used correspond with those from Figure 8. Compared with the 'Truncated Filtering Analysis' there is some loss of accuracy and frequency resolution. However, this analysis method can be automated rather easily.

To check if steeper filter slopes could improve the method, two alternative filters were investigated: a 4th order bandfilter (simply by using the filter twice, thus equivalent with cascading two identical 2nd order sections) and a Gauss filter. In both cases the 3dB bandwidths were F_0–controlled again. Specifically for the Gauss filter the 'valleys' were deeper. The readability of the formants, however, was not improved because of the curvation of the slopes of the Gauss filter (on a vertical log scale); which may falsely suggest the presence of (weak) formants. The 4th order filter gave no noticable improvements of the readability of the formants whatsoever.

6 Conclusion and discussion

Formant determination on voiced speech signals with low formant–to–fundamental values is generally found to be rather disappointing: the signal simply contains not enough information to reach the desired accuracy. Generally, a filter function can not be estimated properly when the test signals are not suitable. From a perception point of view, the consequence is that this inaccuracy is applicable for the presentation of these kinds of signals as well. In this respect the analysis should not suggest a better resolution than the limit present in the signal itself. The preferred type of formant analysis graph, therefore, should present all spectral information about formants with optimal readability and at the same time suppress F_0 effects as much as possible.

Both methods described are useful in this relation. There is no need to select analyzing parameter values dependent on the signal type: the output can be regarded as an optimal spectral display of all kinds of speech–like sounds (periodic or not). The first method (pitch–synchronous 'Truncated Filtering Analysis') gives the best resolution and accuracy. Its output can be regarded as being the cross–correlation of the signal with a truncated damped sinusoid, as a function of its center frequency, which basically seems the target. The requirement to find the exact period crossing, i.e. the position of the closing of the vocal folds, however, is the main drawback as this is difficult to find and to automate, especially in cases with high F_0 and low formant frequencies.

The second method ('Pitch–controlled Bandpass Filter Analysis') can be easily automated and, while sacrificing some frequency resolution and accuracy, presents rather reliable spectral graphs as a basis for formant estimation (for example by automatic peak picking algorithms). Using the local pitch data it is possible to recirculate one local period of the speech sound (the presented 'Praat' script is organized as such). In this way a per–period formant analysis can be performed, which avoids inaccuracy caused by averaging formant shifts within longer intervals, and offers optimal accuracy when measuring formant transients. Of course, it remains possible to window a (steady) part of a speech signal which gives more noise independency but averages possible formant shifts.

The presented 'Praat' scripts are not optimized for speed or efficiency: they merely serve for testing the analysis methods.

Bibliography

Boersma, P. P. G. & D. J. M. Weenink, (1996): *Praat, a system for doing phonetics by computer, version 3.4*, report **132**, Institute of Phonetic Sciences University of Amsterdam [up–to–date version of program and manual downloadable at <http://www.praat.org>].

Lynn, P.A. (1987): *An Introduction to the Analysis and Processing of Signals* (2nd edition), Macmillan Education Ltd.

Randall, R.B. & B.Tech, (1977): "Application of B&K Equipment to Frequency Analysis", Brüel & Kjær, Denmark.
Wempe, T. (1979): "An experimental segment spectrograph based on some notes on frequency analysis of speech segments", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **5**: 44–102.

Appendix A

'Truncated Filtering Analysis' script for the program 'Praat'

```
# Finite components spectral analyzer
# The F0 period of a voiced speech sound must be isolated and selected in advance
# Unvoiced intervals can be chosen freely
# The output is an intensity object with scaled axes for spectral interpretation

form Finite Components Spectrum
  positive Filter_Width_/F0_(Hz) 1/3
  real Lowest_Frequency_(Hz) 0
  positive Highest_Frequency_(Hz) 4000
  positive Dynamic_Range_(dB) 40
endform

Copy... segment

# Resample, if necessary, to 44100
sr = Get sample rate
if sr <> 44100
  Resample... 44100 50
  select Sound segment
  Remove
  select Sound segment_44100
  Rename... segment
endif

d = Get duration
fbw = 'Filter_Width_/F0' /d
fstep = fbw / 2
numsteps = ('Highest_Frequency' – 'Lowest_Frequency') / fstep

# Extend with zeroes (0.4 s) to do FFT on 'isolated' period
Create Sound... embed 0 0.4 44100 0
Formula... self + Sound_segment[col]

To Spectrum
# Create frame for spectrum filter
Copy... filter
# Create initial multiplied spectrum
Copy... mult
Formula... 0

# Create initial accumulated sound with half first step duration
Create Sound... accu 0 'd'/2 44100 0

for i from 1 to numsteps+1
# adjust measuring filter
```

```
   freq =  'Lowest_Frequency' + i * fstep
   select Spectrum filter
   Formula... if row = 1 then – x^2 * 'fbw'^2 else ('freq'^2 – x^2) * 'fbw' * x  fi
... / (('freq'^2 – x^2)^2 + x^2 * 'fbw'^2)

# multiply filter spectrum and signal spectrum
   select Spectrum mult
   Formula... if row=1 then Spectrum_embed[1,col]*Spectrum_filter[1,col]
...  – Spectrum_embed[2,col]*Spectrum_filter[2,col] else
... Spectrum_embed[1,col]*Spectrum_filter[2,col]+Spectrum_embed[2,col]
... * Spectrum_filter[1,col] fi

   To Sound

# concatenate new "filter output" (truncated!) and accumulated "filter outputs"
   select Sound segment
   Copy... filseg
   Formula... Sound_mult[col]
   plus Sound accu
   Concatenate

   select Sound accu
   plus Sound filseg
   plus Sound mult
   Remove
   select Sound chain
   Rename... accu

endfor

beginfreq = 'Lowest_Frequency'
endfreq = 'Highest_Frequency'
drange = 'Dynamic_Range'
select Sound accu
To Intensity... 1/'d' 'd'/3
imax = Get maximum... 0 0 Parabolic
tbegin = 0
tend = numsteps * 'd'

Draw... 'tbegin' 'tend'  'imax'–'drange' 'imax' no
Draw inner box
Axes... 'beginfreq' 'endfreq' 'imax'–'drange' 'imax'
Text bottom... yes Frequency (Hz)
Text left... yes dB/Hz
Marks left every... 1 10 yes yes no
Marks bottom every... 1 1000 yes yes no

select Sound segment
plus Sound embed
plus Sound accu
plus Spectrum embed
plus Spectrum filter
plus Spectrum mult
Remove
```

Remarks

The gradual filter slopes limit the dynamic range, hence the default value of 40 dB.

The script is based on a sampling frequency of 44.1 kHz which makes it possible to listen to sounds via the sound card. Of course, any high value will do.

Although the filtered energy can be estimated directly in 'Praat', the filter sound responses are used in order to be able to apply the Intensity analysis which avarages the per–step energy fluctuations.

Scripts are downloadable from <http://www.fon.hum.uva.nl/wempe>.

Appendix B

F_0–controlled bandfilter analysis script for the program 'Praat'

```
# One F0 period of a voiced speech sound must be isolated and selected in advance;
#           the exact period crossing need not be determined
# Unvoiced intervals can be chosen freely
# The output is an intensity object with scaled axes for spectral interpretation

form 2nd order BF Spectrum
   positive Filter_Width_/F0_(Hz) 1.25
   real Lowest_Frequency_(Hz) 0
   positive Highest_Frequency_(Hz) 4000
   positive Dynamic_Range_(dB) 40
endform

Copy... segment

# Resample, if necessary, to 44100
sr = Get sample rate
if sr <> 44100
   Resample... 44100 50
   select Sound segment
   Remove
   select Sound segment_44100
   Rename... segment
endif

d = Get duration
ns = Get number of samples
fbw = 'Filter_Width_/F0' * 1/d
fstep = fbw / 6
numsteps = ('Highest_Frequency' – 'Lowest_Frequency') / fstep

# Fill 0.1 s with periods
Create Sound... sustained 0 0.1 44100 0
Formula... Sound_segment(x mod 'd')

To Spectrum

# Create frame for spectrum filter
Copy... filter

# Create initial multiplied spectrum
Copy... mult
Formula... 0

# Create initial accumulated sound with half first step duration
```

```
Create Sound... accu 0 0.05 44100 0

for i from 1 to numsteps+1
# adjust measuring filter
   freq =  'Lowest_Frequency' + i * fstep
   select Spectrum filter
   Formula... if row = 1 then ('freq'^2 – x^2) * 'fbw' * x else x^2 * 'fbw'^2  fi
... / (('freq'^2 – x^2)^2 + x^2 * 'fbw'^2)

# multiply filter spectrum and signal spectrum
   select Spectrum mult
   Formula... if row=1 then Spectrum_sustained[1,col]*Spectrum_filter[1,col]
...  – Spectrum_sustained[2,col]*Spectrum_filter[2,col] else
... Spectrum_sustained[1,col]*Spectrum_filter[2,col]+Spectrum_sustained[2,col]
... * Spectrum_filter[1,col] fi

   To Sound

# concatenate new "filter output" and accumulated "filter outputs"
#   limit length of sound mult to length of sound sustained
   select Sound sustained
   Copy... filseg
   Formula... Sound_mult[col]
   plus Sound accu
   Concatenate

   select Sound accu
   plus Sound filseg
   plus Sound mult
   Remove
   select Sound chain
   Rename... accu

endfor

beginfreq = 'Lowest_Frequency'
endfreq = 'Highest_Frequency'
drange = 'Dynamic_Range'
select Sound accu
To Intensity... 1/0.1  0.1/3

select Sound segment
plus Sound sustained
plus Spectrum sustained
plus Spectrum mult
plus Spectrum filter
#plus Sound accu
Remove

select Intensity accu
imax = Get maximum... 0 0 Parabolic
tbegin = 0
tend = numsteps * 0.1
Draw... 'tbegin' 'tend'  'imax'–'drange' 'imax' no
Draw inner box
Axes... 'beginfreq' 'endfreq' 'imax'–'drange' 'imax'

Marks bottom every... 1 1000 yes yes no
Marks left every... 1 10  yes yes no
Text bottom... yes Frequency (Hz)
```

Text left... yes dB/Hz