Time-Domain Formant-Wave-Function Synthesis
Author(s): Xavier Rodet
Source: *Computer Music Journal,* Vol. 8, No. 3 (Autumn, 1984), pp. 9-14
Published by: The MIT Press
Stable URL: http://www.jstor.org/stable/3679809
Accessed: 09-01-2018 05:12 UTC

Xavier Rodet
IRCAM
31, rue Saint-Merri
F-75004 Paris, France

Time-Domain Formant-Wave-Function Synthesis

Introduction

Formant-wave-function (FOF) synthesis is a method for directly calculating the amplitude of the waveform of a signal as a function of time. Many signals can be modeled as a pair: an excitation function followed by a parallel filter. In the FOF method, this pair is replaced by a unique formula describing in a more or less approximate way the output of the filter. Two advantages have motivated the first uses of the FOF technique. On the one hand, in certain cases, the FOF formula can be simplified to a point where calculations are fast and easy (Rodet and Santamarina 1975; Bourgenot and Dechaux 1975; Baumwolspinner 1978; Rodet and Delatre 1979). On the other hand, the FOF method allows the modeling of signals without a need to separate "a priori" the excitation function and the filter (Rodet 1977; Rodet and Delatre 1979).

In formant-wave-function synthesis, the term "formant" is meant in a broad sense. It designates a part of the spectrum (respectively of the signal) that is considered as a whole for a given application. (See the later section entitled "Example of Decomposition of a Signal into Partial Formant Wave Functions.")

Theory

Production of different kinds of signals (speech or instrumental sounds, for instance) can be represented in the form of an excitation function $e(k)$ (with z transform $e(z)$) and a filter, the transfer function of which is $H(z)$ (Fig. 1). The response of the filter to the excitation $e(z)$ is

$$s(z) = e(z) \cdot H(z).$$

This article is reprinted from J. C. Simon, ed., 1980, *Spoken Language Generation and Understanding*, D. Reidel Publishing Company, Dordrecht, Holland.

Suppose that the filter H is linear. If the excitation is a succession of impulses or arches $E(k) = \Sigma e_n(k)$ (where n indexes successive arches), then the response of the filter can be calculated easily as the sum of the responses $s_n(k)$ offset by one period of the fundamental $\tau_0 = 1/F_0$ (F_0 being the fundamental frequency of the excitation and of the response).

In general, $H(z)$ is considered to be a set of parallel filters $H(z) = \Sigma_{i=1}^{q} h_i(z)$ and the response $s_n(k)$ is then the sum of q partial responses $s_n(k) = \Sigma_{i=1}^{q} s_{in}(k)$.

Finally the waveform at period number n is the sum of q partial formant wave functions $s_{in}(k)$, each modeling a formant or a more-or-less broad portion of a given spectrum so as to obtain the best fit to this spectrum or to its most important characteristics.

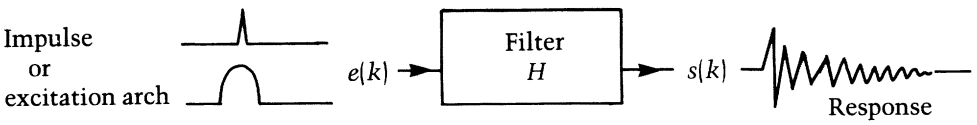
It is impossible to consider the responses $s_n(k)$ as infinite. Generally, the quantization of amplitudes limits the nonzero duration of $s_n(k)$ to a finite value. Thus we call δt_n the duration during which $s_n(k)$ is nonzero.

This model holds true even if the fundamental period is not constant, that is if the delay between the successive excitation arches $e_n(k)$ is not a constant τ_0 but a value τ_n which varies with the index n of the excitation arch: in this case, the successive responses to be summed will be delayed by τ_n respectively.

The same is true if the shape of the excitation arch $e_n(k)$ varies with time: to each arch corresponds a particular response $s_n(k)$ which has to be calculated. Particularly, if the arches of the excitation signal differ solely by a multiplicative coefficient Q_n, the total response is the sum of the responses: $\Sigma Q_n \cdot s_n(k)$.

Finally, in the case of the filter varying with time, these variations have to be taken into account in the formula defining $s_n(k)$. However, if the variations of the filter during one fundamental period are small, one can consider the characteristics of the filter as constant during the time interval δt_n during which the response $s_n(k)$ is calculated. But each successive response has to be calculated as a

Fig. 1. Excitation function and filter model of signal production.



function of the characteristics of the filter at the corresponding excitation time.

One of the advantages of the FOF method is that it allows the modeling of a spectrum of complex shape by means of a limited number of functions (as few as one—see the next section). Furthermore, in many cases the cost of calculation on a general purpose computer is considerably less with the FOF technique than with the excitation-function-filter model. It should also be mentioned that the precision of calculations necessary for the FOF method is less than that required for filters: usually it is only the precision required for the final signal (for instance, a 12- to 16-bit integer for a high-fidelity audio signal). Generally, the calculations can be made in fixed point and with tables of limited size (Rodet and Delatre 1979). Finally, the sum of several simple FOFs allows us to model spectra of complex shapes with great precision (for example those of vocal and instrumental sounds). It is notable that the parameters of the model are particularly representative from a perceptual point of view.

Example of Decomposition of a Signal into Partial Formant Wave Functions

This method allows one to separate a signal $P(k)$ into a number q of partial FOFs corresponding to q regions Φ_i on the frequency axis (Rodet and Delatre 1979; Rodet, Delatre, and Razzam 1979). Those regions are distinct but any given region is not necessarily contiguous. The chosen portion of signal $P(k)$ corresponds exactly to one fundamental period. The spectrum of the signal is modeled in the form of an m-pole linear predictive filter ($G/A(z)$). The parallel structure of this filter,

$$\sum_{l=1}^{m/2} \frac{c_l + d_l z^{-1}}{1 + a_l z^{-1} + b_l z^{-2}},$$

can be decomposed into q groups respectively corresponding to the q regions Φ_i: Each group γ_i is the sum of those sections $(c_l + d_l z^{-1})/(1 + a_l z^{-1} + b_l z^{-2})$ the poles of whose frequencies are contained in the corresponding region Φ_i. Filtering of the prediction error $E(k)$ through each of the q groups γ_i gives q partial wave functions $p_i(k)$. Thus they are the responses of the q parallel groups γ_i to one arch of the excitation signal with the property that their sum is equal to the original signal to be modeled

$$\sum_{i=1}^{q} p_i(k) = P(k).$$

Thus, these partial wave functions $p(k)$ can be considered as FOFs and, after being tabulated can be used according to the method described previously. The presence of zeros in the transfer function is naturally taken into account by the model. Each of these FOF functions can independently be modified in amplitude or in time so as to change the spectral shape or so as to displace a "formant" region Φ_i. If it is not necessary to modify these FOFs independently, their number can be reduced to 1 ($q = 1$). This allows the synthesis of a sound with a very complex spectral shape by means of a single table lookup. One can still change the fundamental frequency or apply a homothetic transform to the spectrum.

An Applied Example

Certain models (parallel formant speech synthesizers for instance) utilize a set of second-order sections in parallel. Consider such a section, the transfer function of which is:

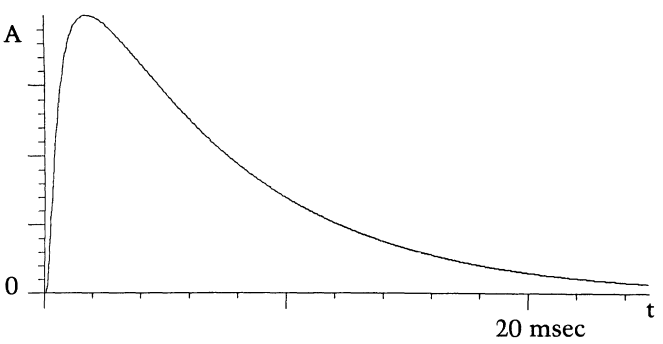
$$H(z) = \frac{c + d\,z^{-1}}{1 + a\,z^{-1} + b\,z^{-2}} = \frac{c + d\,z^{-1}}{(1 - p\,z^{-1})(1 - p^*\,z^{-1})}.$$

Its impulse response is the formant wave function

$$s(k) = G\,e^{-\alpha k} \sin(\omega k + \phi)$$

where

Fig. 2. Envelope $Ae^{-\alpha}/[(t - c)^2/t]$ *where* $\alpha/\pi = 50$ *Hz and* $c = 1.8$ *msec.*



$\alpha = -\log(P)$;

$\omega = \text{Arg}(P)$; and

$$G = \frac{c}{\sin(\phi)}.$$

Thus, some projects have made use of a sinusoidal wave multiplied by an exponentially damped amplitude envelope.

For natural sounds the excitation function is not a unit impulse. To obtain a more precise control of the spectrum, the function $e^{-\alpha}[(t - c)^{-2}/t]$ (Fig. 2) has been used as an amplitude envelope (Rodet and Delatre 1979; Rodet, Delatre, and Razzan 1979). Vocal sounds of very high quality have been obtained this way.

For modeling arbitrary spectra one can use other types of envelopes $A(t)$ chosen according to their spectra.

In effect, the spectrum of the FOF so obtained $A(t) \cdot \sin(\omega_c t + \phi)$, is that of the envelope translated to the center frequency $\omega c/2\pi$. In particular, the windows used in spectral analysis are interesting envelopes (Harris 1978). They can be chosen for two reasons:

- *to concentrate the energy* in a narrow frequency band and during a temporal window as small as possible (in order to reduce the calculations and obtain a close fit to formant values during transitions) and
- *to approximate a portion of the spectrum of any shape* with a principal peak and lateral lobes of greater or lesser magnitude.

Second Example

The following FOF is preferred for its interesting properties:

$$s(k) = 0 \quad \text{for } k \leq 0$$

$$s(k) = \frac{1}{2}(1 - \cos[\beta k])e^{-\alpha k} \sin(\omega_c \cdot k + \phi) \quad \text{for}$$

$$0 \leq k \leq k_1, \text{ with } k_1 = \frac{\pi}{\beta}$$

$$s(k) = e^{-\alpha k} \sin(\omega_c \cdot k + \phi) \quad \text{for } k \geq k_1$$

It is again a sinusoid function shaped by an amplitude envelope $A(k)$. This envelope is a damped exponential, the initial discontinuity of which is smoothed by multiplication by $(1/2)(1 - \cos[\beta k])$ for a duration of k_1 samples so that $\beta k_1 = \pi$ (Fig. 3).

One obtains thus an amplitude envelope $A(k)$ which has an attack for a duration of approximately k_1 samples and a $e^{-\alpha k}$ decay after the attack. This amplitude envelope $A(k)$ presents no first- or second-order discontinuity and is obtained very simply by table lookup for $(1/2)(1 - \cos[\beta k])$ and $\sin(\omega k)$ and by successive multiplications by $e^{-\alpha}$ for $e^{-\alpha k}$.

The Fourier transform of this envelope is

$$\psi(\omega) = \int_{-\infty}^{+\infty} A(t)e^{-i\omega t}dt$$

$$= \int_0^{\frac{\pi}{\beta}} \frac{1}{2}(1 - \cos[\beta t])\, e^{-(\alpha + i\omega)t}dt + \int_{\frac{\pi}{\beta}}^{+\infty} e^{-(\alpha + i\omega)t}dt.$$

Thus

$$\psi(\omega) = \frac{\beta^2}{2} \frac{e^{-(\alpha + i\omega)\frac{\pi}{\beta}} + 1}{(\alpha + i\omega)((\alpha + i\omega)^2 + \beta^2)},$$

the absolute value of which is

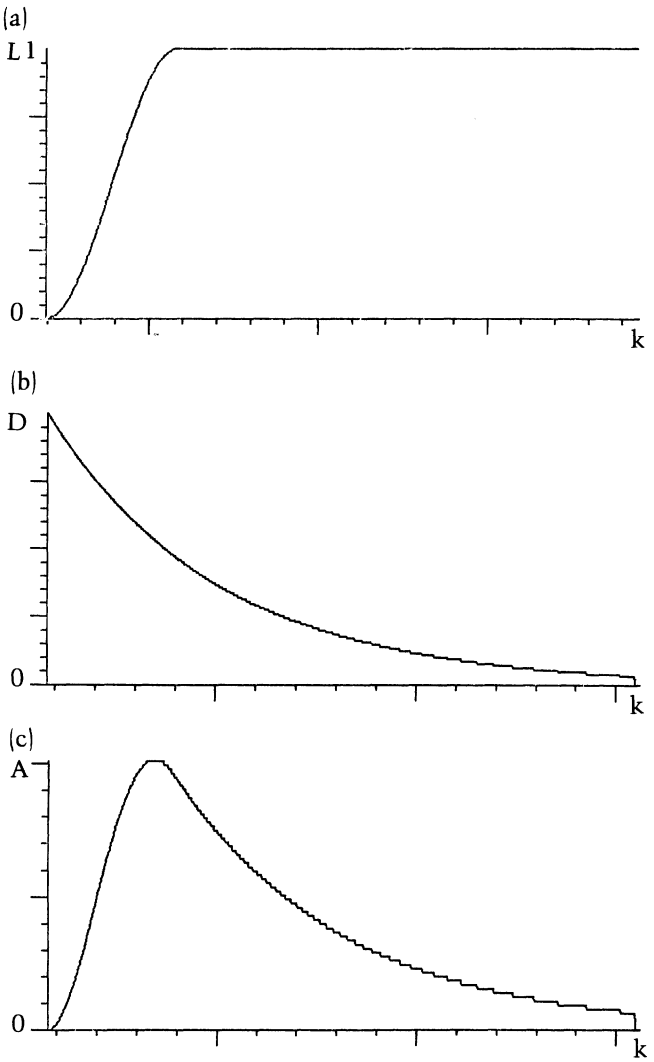
$$|\psi(\omega)| = \frac{\beta^2}{2} \frac{\sqrt{M^2 + 1 + 2M\cos\left(\omega \frac{\pi}{\beta}\right)}}{\sqrt{\alpha^2 + \omega^2}\sqrt{(\alpha^2 + \omega^2)^2 + \beta^2(\beta^2 + 2\alpha^2 - 2\omega^2)}} \quad (1)$$

with $M = e^{-(\alpha\pi/\beta)}$.

For speech synthesis, the orders of magnitude of the values are

$\alpha = (\text{bandwidth}) \cdot \pi \simeq 10^2\pi$ Hz

Fig. 3. In (a) is shown the attack of envelope A, (b) is the decay, and (c) is the combined envelope A(k) = L(k) · D(k).



$$\beta = \frac{\pi}{\text{attack duration}} \simeq 10^3\pi \text{ Hz}.$$

Thus one can consider as a first approximation $\beta^2 \gg \alpha^2$.

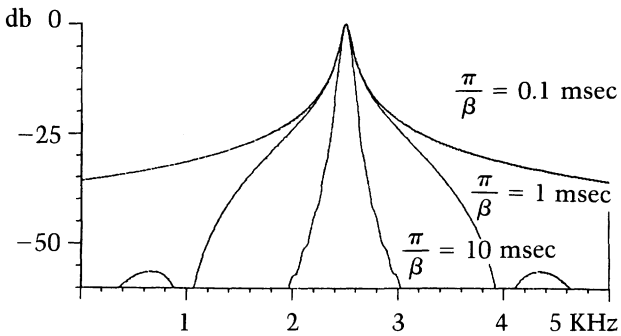
Moreover, if $\omega(\pi/\beta)$ is small with respect to $\pi/2$ for frequencies close to the center frequency, $F_c = \omega c/2\pi$, one can consider $\cos(\omega[\pi/\beta]) \simeq 1$ and $\beta^2 > \omega^2$.

It follows that

$$|\psi(\omega)| \simeq \frac{\beta^2}{2} \frac{\sqrt{M^2 + 2M + 1}}{\sqrt{\alpha^2 + \omega^2}\sqrt{\beta^4}} = \frac{M+1}{2} \frac{1}{\sqrt{\alpha^2 + \omega^2}}.$$

Thus, the shape of the power spectrum of our FOF

Fig. 4. Power spectrum of the FOF. A(k) sin($\omega_c \cdot k + \phi$) for different values of π/β, where ω_c = 2.5 KHz and α/π = 80 Hz.



$s(k)$ is of the form $K/[\alpha^2 + (\omega_c - \omega)^2]$ in the neighborhood of the center frequency $\omega_c/2\pi$. It is independent of β and nearly identical to the transfer function of a second-order filter section, the central pulsation of which is ω_c and the bandwidth is α/π. Thus the parameter α controls the −6 db bandwidth of the spectrum of this "formant." The parameter β controls the width of the "skirts" of the formant peak. It does not modify the bandwidth at −6 db, which is a quite remarkable property (Fig. 4).

Finally, according to equation (1), for a given α and β, it is easy to normalize the amplitude of the resulting spectrum. A parallel FOF synthesizer program has been written in software. (See "The CHANT Project" by Rodet, Potard, and Barrière elsewhere in this issue.) Its structure is shown in Fig. 5. For formant number i

ω_i = center pulsation;

A_i = amplitude;

BW_i = bandwidth; and

$\frac{\pi}{\beta_i}$ = width of the skirts.

This program (CHANT) is used for synthesis of singing voices and of instrumental sounds. The different steps of the modeling include semiautomatic analysis of the spectrum of a given sound and extraction of gross formant characteristics ($\{\omega_i, BW_i, A_i, \beta_i\}$ for $1 \leq i \leq q$) and fundamental frequency (Fig. 6).

With CHANT, it is also possible to adjust the parameters so as to model the natural spectrum as closely as possible (Fig. 7). The parameters are shown in Table 1.

Fig. 5. Structure of a parallel FOF synthesizer.

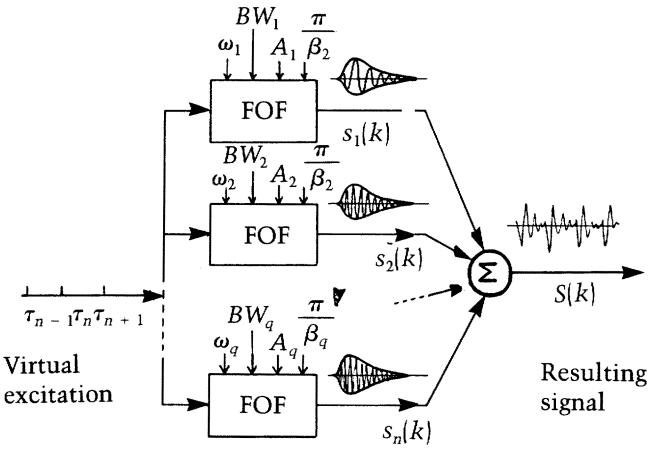


Fig. 6. Comparison of a natural voice spectrum with the envelope of a synthetic spectrum (outline) calculated according to Eq. (1), from a gross estimation of the parameters.

Fig. 6

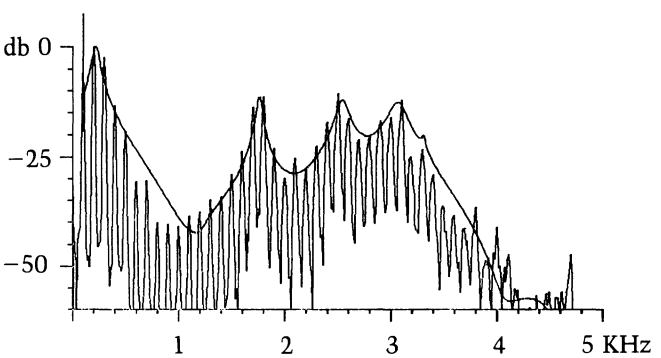


Fig. 7

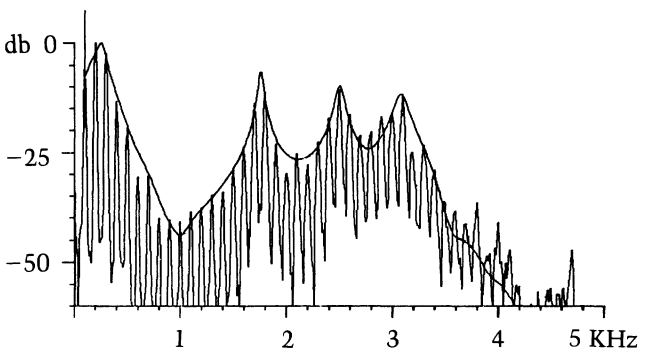


Fig. 7. Adjustment of parameters so as to model the natural vocal spectrum as closely as possible. The parameters are listed in Table 1.

Fig. 8

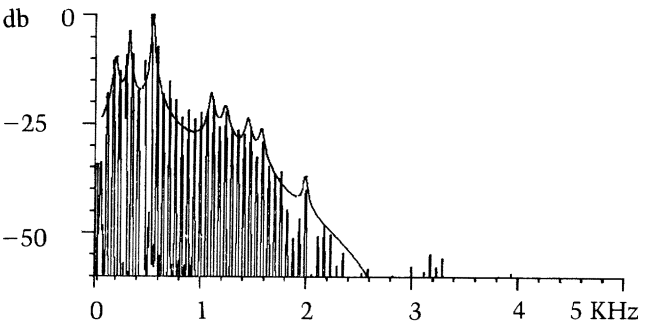
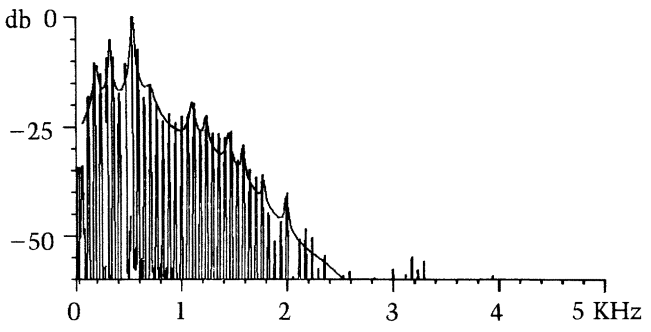


Fig. 8. Comparison of a natural bassoon spectrum with the envelope of a synthetic spectrum (outline) calculated according to a gross estimation of the parameters.

Table 1. Parameters for modeling a voice spectrum

$\frac{\pi}{\beta_1}$ = .002 sec	F_1 = 260 Hz A_1 = .029 BW_1 = 70 Hz pu = .002 sec
$\frac{\pi}{\beta_2}$ = .0015 sec	F_2 = 1764 Hz A_2 = .021 BW_2 = 45 Hz pu = .0015 sec
$\frac{\pi}{\beta_3}$ = .0015 sec	F_3 = 2510 Hz A_3 = .0146 BW_3 = 80 Hz pu = .0015 sec
$\frac{\pi}{\beta_4}$ = .003 sec	F_4 = 3090 Hz A_4 = .011 BW_4 = 130 Hz pu = .003 sec
$\frac{\pi}{\beta_5}$ = .001 sec	F_5 = 3310 Hz A_5 = .00061 BW_5 = 150 Hz pu = .001 sec

Fig. 9. Adjustment of parameters so as to model a bassoon spectrum as closely as possible.



Figures 8 and 9 show the results of the same operations on the spectrum of a bassoon.

Conclusion

The parallel formant-wave-function method has been used for a number of years for economy of computation in speech synthesis. It has been shown that the development of adequate formulas produces a close model to the excitation-filter model. The FOF model allows synthesis of vocal and instrumental sounds of excellent quality on general-purpose computers with limited calculation cost, and without calculations of great precision.

The parameters of our model include the fundamental frequency, and for each formant, the center frequency, the bandwidth at −6 db, the width of the skirts, and the amplitude. These parameters are particularly easy to determine and to adjust, and they are especially relevant from a perceptual point of view.

Provided that the parameters are correctly chosen, a very high sound quality can be obtained in the synthesis. Thus a wide variety of sounds can be synthesized, from percussion instruments to voice, from reed instruments to strings.

References

Baumwolspinner, M. 1978. "Speech Generation through Waveform Synthesis." *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. New York: IEEE, pp. 179–182.

Bourgenot, J. S., and C. Dechaux. 1975. "Codage de la parole à faible débit: le vocoder CIPHON." *Revue Technique Thompson-CSF* 7(4):13–17.

Harris, F. 1978. "On the Use of Windows for Harmonic Analysis with Discrete Fourier Transforms." *Proceedings of the IEEE* 66(1):51–83.

Rodet, X. 1977. "Analyse du signal vocale dans sa représentation amplitude-temps: Synthèse de la parole par règles." Thèse d'Etat. Paris: Université Paris VI.

Rodet, X., and C. Santamarina. 1975. "Synthèse, sur un miniordinateur, du signal vocale dans sa représentation amplitude-temps." *Actes des sixième journées d'étude sur la parole du GALF, Toulouse*. Paris: GALF, pp. 364–371.

Rodet, X., and J. Delatre. 1979. "Time-domain Speech Synthesis by Rules Using a Flexible and Fast Signal Management System." *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Washington, D.C., 2–4 April*. New York: IEEE, pp. 895–898.

Rodet, X., J. Delatre, and M. Razzam. 1979. "Construction du signal vocale dans le domaine temporel." *Actes des dixième journées d'étude sur la parole du GALF, Grenoble*. Paris: GALF, pp. 80–88.