

## SPEECH ANALYSIS AND SYNTHESIS METHODS BASED ON SPECTRAL ENVELOPES AND VOICED/UNVOICED FUNCTIONS

X. Rodet<sup>(1)(2)</sup>, P. Depalle<sup>(1)(3)</sup> & G. Poirot<sup>(1)</sup>

### ABSTRACT

The Institut de Recherche et de Communication Acoustique/Musique (IRCAM) is involved in sound analysis and synthesis for contemporary Music Creation. Accurate analysis and high quality synthesis of sounds are thus extensively studied. We describe here some of the methods that have been developed for speech and singing voice. They can be used for accurate analysis, for sound processing (such as time or spectral warping, filtering, pitch transposition, etc...) and for synthesis-by-rule. The goal is also to provide powerful tools and to get sounds with as high a quality as possible since they will be used in pieces and concerts, but speech applications are equally in view <sup>(4)</sup>. At present, speech signals are digitized at 16KHz. on 16 bits. Higher sampling rates are considered.

### SPECTRAL ESTIMATION

We use the classical source-filter approximation to model speech production. Consequently, on one side we look for an estimation of the vocal tract, and on the other side we need a parametric representation of the excitation in terms of fundamental frequency, source spectrum slope and voiced/unvoiced characteristics.

We have chosen to model the vocal tract with an AutoRegressive (AR) model because a good estimation can then be obtained if the number of poles is high enough. In addition resolution methods are simple and efficient. Since the signal is non-stationary, an adaptive method has been selected. We use the recursive adaptive lattice LPC which was proposed by [Viswanathan 78]. In addition, the  $K_1$  reflection coefficients are guaranteed to be less than one in magnitude so that the synthesis filter is stable.

To accurately model fast transitions, such as in consonants, it is necessary that the whitening filter of the analysis adapt itself as fast as possible. Classically, an exponential sliding window is applied on the error signal. The value of the exponential decay coefficient is usually chosen close to 1 (typically .995 at 16 KHz), but we can sometimes observe, especially when the energy of the signal is abruptly attenuated (for instance in occlusives) that the filter tends to maintain characteristics of the past, so that the synthesis exhibits a kind of "reverberated" quality. Inversely, if the exponential decay coefficient is too small, the optimisation criterion does not remain valid and the spectral envelope is not representative of the Power Spectral Density of the signal.

Thus we use a window with better properties for the analysis, i.e. one which is rather flat on the right end and smoothly damped to zero on the left end. This window is applied, either to the sound signal itself, or to the error signal. In the first case, the windowed signal is analysed by the pre-cited method. In the second case, the analysis-method is itself modified since the optimisation criterion is modified. As an example, in figure 1 the left-half of a 3 Coefficient Blackman-Harris window [Harris 1978] has been applied to the source signal. It leads to an extremely accurate estimation, even on a segment as short as 20 ms.

### VOICED/UNVOICED DECISION

It is common to segment speech signals into voiced and unvoiced regions. However this decision is difficult and leads to "errors" which have a dramatic effect on synthesis quality. Furthermore, many sounds, such as voiced fricatives, exhibit a short-time spectrum which is partly harmonic (voiced), partly random (unvoiced). Finally, LPC synthesis is often heard as

(1) IRCAM, 31 rue Saint Merri, 75004 Paris, France, Tel.: (1) 42 77 12 33 ext. 48-27, 48-14, uucp net address: decvax!seismo!ircam!rod,phd,gip

(2) Also with L.A.F.O.R.I.A., Université Paris-6, France.

(3) Also with Ecole Supérieure d'Electricité, Metz, France.

(4) This research was partly supported by CNET-Lannion, France.

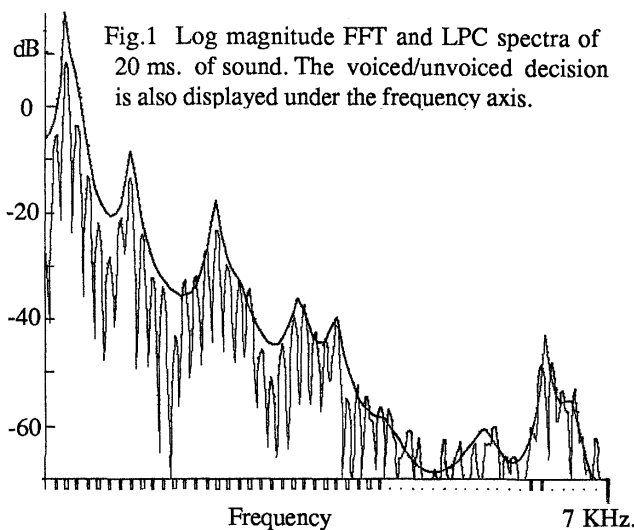


Fig.1 Log magnitude FFT and LPC spectra of 20 ms. of sound. The voiced/unvoiced decision is also displayed under the frequency axis.

"buzzy", this being due, in part, to the pulse excitation which is "too periodic": even in vowels, the short time spectrum is often not purely harmonic above 3.5 or 4 kHz. Among other reasons, the linear approximation of the vocal tract is no longer valid at such frequencies. We have developed an algorithm giving a voiced/unvoiced decision in frequency bands covering the entire spectrum. This method is analogous to the one proposed by [Griffin 85] but presents several improvements. In particular, since the decision concerns the excitation, we take it on the LPC residual.

Let  $f_0, f_1, \dots, f_k, \dots$  be the harmonic frequencies looked for in the residual spectrum  $S_r(f)$ , and let  $S_h(f)$  be a spectrum reconstructed to be exactly harmonic according to the fundamental frequency and according to the analysis window, and with amplitudes found in  $S_r(f)$ . For each possible harmonic, that is for each  $k$ , we compute

$$N(k; f_0) = \int_{f_k - f_0/2}^{f_k + f_0/2} (|S_r(f)| - |S_h(f)|)^2 df \quad (1)$$

$$\xi(k; f_0) = 1 - \frac{N(k; f_0)}{\max N(k; f_0)} \quad (2)$$

where  $\max N(k; f_0)$  is the maximum value that  $N(k; f_0)$  can take for any  $S_r(f)$  values in the band  $k$ . Thus  $\xi(k; f_0)$  takes value in  $[0,1]$ , being 1 only for exactly harmonic sinusoids.

The voiced/unvoiced decision is then taken by comparing  $\xi(k; f_0)$  to a threshold (typically 0.8). An example of the decision is shown in Fig.1 under the frequency-axis: each harmonic is represented by a short vertical double line and unvoiced bands by a dot. The results of this algorithm have been found to agree well with decisions made by visual examination of signals and spectra. When used in synthesis, the voiced/unvoiced decision in harmonic frequency bands leads to a remarkable improvement in the "natural" quality.

#### CALCULATION OF THE GAIN FACTOR

Modeling the signal with an AR model, gives us a spectral envelope value at normalised frequency  $f$ :

$$G^2 \cdot \frac{1}{|A(e^{2\pi jf})|} \quad (3)$$

But it must be noted that the whitening of the speech signal is not perfect, whereas the synthesis is done with a flat spectrum excitation. We thus calculate the gain factor in order to match the energy of the natural and synthetic signals.

Let  $X(z)$  be the  $z$ -transform of one period of the original signal, and let us equalise the magnitude spectra of both the original and the synthetic signals at each normalised frequency  $f=k/n$  where  $n$  is the number of samples of the fundamental period:

$$|X(e^{2\pi jf})| = G^2 \cdot \frac{1}{|A(e^{2\pi jf})|^2}, \forall f = k/n \quad (4)$$

If we sum all such equalities for  $f=k/n$ :

$$\sum_{k=0}^{n-1} |X(e^{2\pi jf})|^2 = G^2 \cdot \sum_{k=0}^{n-1} \frac{1}{|A(e^{2\pi jf})|^2} \quad (5)$$

Then, using Parseval's theorem:

$$\frac{1}{n} \sum_{k=0}^{n-1} |X(e^{2\pi jf})|^2 = \sum_{k=0}^{n-1} |x_k|^2 \quad (6)$$

where the  $x_k$  are the samples of the residual.

Finally:

$$G^2 = n \cdot \frac{\sum_{k=0}^{n-1} |x_k|^2}{\sum_{k=0}^{n-1} \frac{1}{|A(e^{2\pi jf})|^2}} \quad (7)$$

## FORMANT CODING OF THE SPECTRUM ENVELOPPE

The maxima of the spectrum envelope are known to convey the main part of the perceptual information concerning the vocal tract. So we code the spectrum envelope in terms of its maxima (abusively called here formants). They are characterised by their central frequencies, their amplitudes and their 3-dB bandwidths.

The method developed at University of Paris-6 [Montacie 87] uses the Bairstow algorithm to find the roots of  $A(z)$ . At each step, the initial value for this algorithm is an estimation of a pole computed from the absolute maximum found on the spectral envelope. Finally the roots of  $A(z)$  are separated into two sets: the first receives real poles, the global contribution of which represents the source spectrum slope. The  $p$  pole-pairs found in the second set are sorted in  $m$  classes ( $m \leq p$ ) corresponding to  $m$  maxima, since several pole-pairs may be grouped to contribute to the same maximum, that is to a unique "formant".

## SOURCE-FILTER SYNTHESIS

In the synthesis step, the signal must be reconstructed so as to respect the LPC spectral envelope and the voiced/unvoiced decision (including  $f_0$ ). Precisely, the synthetic signal must be a sum of harmonic partials (for those bands found voiced) with amplitudes according to the spectral envelope, and of a noise the spectral density of which is given by the spectral envelope. Several methods may be used for signal reconstruction.

The source is chosen with a flat spectrum envelope. We built it as the sum of a filtered white noise (unvoiced excitation) and of a filtered quasi-periodic train of pulses with the fundamental frequency found during the analysis or any other function for a different intonation.

The unvoiced excitation is obtained by frequency domain (FFT) filtering of a white noise according to the voiced/unvoiced excitation: that is, the filter transfer function is 1. for the unvoiced frequency bands and 0. for the voiced ones. The complementary transfer function (1. for voiced bands, 0. for unvoiced ones) is used to filter a quasi-periodic train of pulses. The pulse is a flat spectrum one (windowed  $\sin(x)/x$ ) thus allowing any value for the fundamental frequency (a unit pulse would be limited to fundamental frequencies  $f_0 = SR/n$  where  $SR$  is the sampling rate and  $n$  is an integer value).

The first implementation uses a lattice filter [Poirot 86] with the parameters  $K_i$  found in the LPC analysis. There is one set of parameters per pitch period (typically a hundred samples) since the analysis is pitch synchronous. For a good sound quality, it is necessary to interpolate the parameters between successive analysis values: no difference could be heard between linear interpolation of the  $K_i$  and interpolation of the Log Area Ratio (LAR) parameters.

In a second method, the desired spectral shape is applied during the frequency domain filtering: the 1. of the transfert function are simply replaced by the spectral envelope values  $a_k$  at the corresponding frequency  $f=k.f_0$  for the  $k^{\text{th}}$  band centered around frequency  $f$ :

$$a_k = \frac{G}{|A(e^{2\pi j f})|} \quad (8)$$

### ADDITIVE SYNTHESIS

In additive synthesis, the voiced signal is obtained as a sum of sinusoids with harmonic frequencies of the fundamantal, i.e.  $f_0, 2.f_0, \dots, k.f_0, \dots$ , and amplitudes  $a_k$  as above. The unvoiced signal is a sum of narrow-band random signals  $R_k(t)$ , with bandwidth  $f_0$ , obtained by modulation of two orthogonal carriers. Let  $M_k(t)$  and  $N_k(t)$  be low-pass random signals in the band  $[-f_0/2, f_0/2]$ :

$$R_k(t) = a_k(t) \cdot [M_k(t) \cdot \cos(j2\pi k f_0 t) + N_k(t) \cdot \sin(j2\pi k f_0 t)] \quad (9)$$

Formant-Wave-Function synthesis is also being tested [D'Alexandro87]. The best quality was obtained with lattice filtering but additive synthesis is very close to it. Frequency domain filtering (which is analogous to channel vocoder) was found clearly inferior.

Different examples of voice analysis/synthesis, with or without  $f_0$  modification, show that a high quality sound has been obtained, essentially without the usual defaults of LPC like "buzziness".

### APPLICATIONS

The analysis/synthesis methods exposed in this paper have been developed for two main purposes. The first one is processing of voice for musical applications, such as time or frequency warping, changes of voicing characteristics,  $f_0$ , etc...

The second purpose is text-to-speech synthesis-by-rule. Formant-coded segments of speech (or singing voice) such as diphones, will be used in concatenation rules for high quality synthesis-by-rule. The method will benefit from both LPC diphones and formant approaches [Rodet85]. However, other applications where a high quality parametric coding of voice is needed could benefit from our analysis and synthesis methods.

### REFERENCE

[D'Alexandro87] C. D'Alexandro, X. Rodet, Fonctions d'Onde Formantiques, Extraction des Paramètres et Synthèse vocale *to appear in* Actes des 16<sup>èmes</sup> journées d'étude sur la parole de la Société Française d'Acoustique, Hammamet, Tunisie, October 87.

[Harris 78] F. Harris, On the Use of Windows for Harmonic Analysis with Discrete Fourier Transform Proceedings of the IEEE 66(1):51-83.

[Griffin 85] D.W. Griffin, J.S. Lim, A New Model-Based Speech Analysis/Synthesis System IEEE-ICASSP, Tampa, Fl., March 85.

[Montacie 87] C. Montacie, Une détection plus sûre des Formants, rapport interne LAFORIA, Université Paris-6, mars 1987.

[Poirot 86] G. Poirot, P. Willequet, Vocodeur à Prédiction Linéaire, Rapport de DEA, Université du Maine, France, Nov. 1986.

[Rodet 85] X. Rodet, P. Depalle, Synthesis by Rule: LPC Diphones and Calculation of Formant Trajectories, IEEE-ICASSP, Tampa, Fl., March 85.