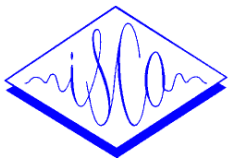

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

A PARAMETRIC MODEL OF SPEECH SIGNALS : APPLICATION TO HIGH QUALITY SPEECH SYNTHESIS BY SPECTRAL AND PROSODIC MODIFICATIONS

Thierry GALAS - Xavier RODET

LAFORIA - UA CNRS N°1095 - PARIS VI - 75252 PARIS CEDEX 05

ABSTRACT

We propose here a new parametric model and its application to speech synthesis. In our source-filter model, the source is described by spectro temporal events. The filter combine an all-pole filter model for the vocal tract and a de-emphasis filter corresponding to the lip radiation and glottal spectrum slope. Source events are singular or belong to a continuum or pseudo-continuum of events. Examples of singular events are the burst of noise at release of plosive or isolated glottal pulses. Pseudo-continua of events are quasi periodic glottal pulses with their intrinsic irregularities with possible superimposed fricative noise, or pure noise signals as in unvoiced fricatives. Our model allows for a precise and perceptually satisfying description of speech signal and simultaneously provides more flexibility for prosodic modifications. We present an analysis method according to our model using any spectral estimation technique such as AR or homomorphic estimations. We also present an overlap-add synthesis method using the analysis data. We show that our method can be interpreted in terms of frequency domain spectral interpolation as an ARMA model.

1. INTRODUCTION.

The use of a source-filter model of speech is commonly justified by some over simplified assumptions:

- phase insensibility of the ear.
- all-pole model of the vocal tract.
- periodic pulse or broad band random noise as the excitation signal.

In consequence the quality of the reconstructed signal with such an analysis-synthesis system is limited (this is the case of the classical LPC vocoder with mutually exclussive pulse and noise excitations). A better description of the LPC residual is obtained with MLPC as proposed by [ATAL 82]. With this method, most of the amplitude and phase characteristics of the residual are reconstructed. However, its quality is still not satisfactory and prosodic modifications require more informations such as pitch period locations. Since then, some semi-parametric methods have been proposed to modify the prosodic characteristics of residuals or speech signals [CHARPENTIER 89]. But these methods, lacking of an explicit production model, suffer from some limitations in the quality and in the range of the possible prosodic modifications. At first, we define our production model. Then we present an analysis method according to model using any spectral estimation technique such as AR or homomorphic estimations. We also present an overlap-add synthesis method using the analysis data. We show that our method can be interpreted in terms of frequency domain spectral interpolation as an ARMA model.

2. PRODUCTION MODEL

In our source-filter model, the source is described by spectro-temporal events. The filter combine an all-pole filter model for the vocal tract and a de-emphasis filter corresponding to the lip radiation and glottal spectrum slope. The all-pole filter is characterized by its parameters (Log Area Ratio, reflection coefficients).

Source events are singular or belong to a continuum or pseudo-continuum of events. Singular events are typically the burst of noise at release of plosive or isolated glottal pulses. ther are two kinds of pseudo-continua of events :

- quasi periodic glottal pulses with their intrinsic irregularities with possible superimposed fricative noise, each glottal pulse is an event of the pseudo continuum.

- pure noise signals as in unvoiced fricatives, wich form a noise continuum.

Singular events are characterized by their locations, and complex spectra. Each event of a pseudo-continuum is characterized by its location, its complex spectrum and a local pseudo-period duration. Noise continuum are only characterized by their magnitude spectra and locations on a centisecond rate basis (for a synthetic vision of the model see Figure 1.).

So our model takes phase spectrum into account. A magnitude spectrum parameter allow to correct unadaptation of an pole filter model for the nasal-vocal system when necessary. A set of parameters for each glottal pulse respect their intrinsic irregularities. Moreover it contains sufficient informations to identify all major acoustics events [Abry 84] such VO (Voice Onset). Then it is possible to use high level knowledge, such as non linear variation of VOT duration along with speech rate modifications.

3. ANALYSIS PROCESS.

For an overview of the process see Figure 2.

3.1 Events Labeling.

The first step of this analysis process is a carefull labelling of the speech signal events. The result is a data sequence $E= \{ E_i = (L_i , T_i) \}$. L_i is the location of the event E_i, and T_i is the type of this event E_i. Events type could be :

- singular events.
- first event of a continuum of pseudo-periodic events.
- pseudo-periodic events (glottal pulses).
- last event of a continuum of pseudo-periodic events.
- beginning of a noise continuum.
- noise continuum (centisecond rate).
- end of a noise continuum.

Events E_i are ordered along their locations L_i. Figure 3 show such a labelling on a signal with a singular event and the beginning of a continuum of pseudo-periodic events (vertical lines are drawed at events locations).

This analysis step is manualy done. As the first application of this analysis-synthesis is automatic text to speech synthesis by diphones concatenation, that does not matter very much. But an automatic labelling is still possible with a greater error rate.

3.2 Envelopes modeling.

This envelopes modeling is synchronized with events locations. All power spectrum estimation methods are useables such as autocorrelation, covariance, Burg-covariance, homomorphic deconvolution, cepstrum true envelope[Imai 79]. Our preferance is given to discret cepstrum[Galas 90] wich performs equally well with men voices and high pitched women voices. The result is a power spectrum $P_i(\omega)$ for each event E_i.

3.4 Inverse Filtering.

Then we search for the all pole filter whose transfert function is the best approximation of the power spectrum envelope resulting of last step. If we choose Itakura & Saito distance [Itakura 68] as error criterion, we have to solve a Yules-Walker system. The matrix coefficients are the envelope autocorrelation coefficients. We compute this coefficients by FFT. We use Durbin recursion to solve the Yules-Walker system. The stability of the resulting all-pole filter is assured. This is a method to stabilise covariance result without poles extraction. For each event E_i.the result is an all-pole filter with a transfert function $1/A_i(z)$.

The inverse filtering is performed after signal pre-emphasis. Filter reflection coefficients are linearly interpolated. Log Area Ratio could be most adapted.

3.5 Singular Events Analysis.

After product of the signal by a window centered on the event location (typically a Hamming of 20-25ms duration), the complex spectrum is evaluate using a FFT. The result is a complex spectrum $C_i(\omega)$

3.6 Pseudo-Periodic Events Analysis.

After product of the signal by a window centered on the event location (typically a Hamming of three pseudo period duration), the complex spectrum is evaluate using a FFT. After a pitch detection, harmonic partials are located on this spectrum. We note F_i the local fondamental frequency.

The values of log magnitude and phase spectra at harmonics locations are linearly interpolated to obtain continuous magnitude and phase spectra. Phase is interpolated with a minimun phase excursion constraint. The result is a complex spectrum $C_i(\omega)$

3.7 Noise Continua Analysis.

Like singular events analysis, after product of the signal by a window centered on the event (typically a Hamming of 20-25ms duration), the complex spectrum is evaluate using a FFT. The result is a complex spectrum $C_i(\omega)$.

3.8 Analysis Parameters.

The analysis result is an events parameters sequence P defined by :

$P = \{ P_i = (E_i , A_i, C_i, F_i) \}$.

E_i is the event of location L_i and T_i is the type T_i , A_i are parameters of an all-pole filter, C_i is the complex spectrum resulting from residual analysis, F_i is the local fondamental frequency, null in case of a noise or singular event .

4. SYNTHESIS PROCESS

4.1 Singular Events Synthesis.

We synthetise these events by a simple inverse Fourier transform of the complex spectrum parameter.C_i.They are located at the exact location parameter.

4.2 Noise Continuum Synthesis.

Noise continuum result from the FFT filtering of a white noise N, according to the magnitude spectrum parameter $\|C_i\|$.

4.3 Quasi-Periodic Events Synthesis.

New locations L_j are computed according to the F_i parameter. Modification of F_i parameter could involve events insertion or deletion. C_j parameter result from C_i parameter interpolation (nearest neighbour). Like singular events, we synthetise these events by a simple inverse Fourier transform of the complex spectrum parameter C_j.

4.4 Residual Construction by Events Overlap-Add.

Each event is lagged by a window W_i centered on the event location. This window is typically a Hamming of 20-25 ms in case of singular events. For pseudo-periodic events the window duration is 2 to 3 periods duration. These windowed events are then overlapped and added. Each sample is normalized such that the sum of windows coefficients would be constant. In fact, it is only a linear interpolation between events. The resulting signal is then added with the noise

continua. Given FT the Fourier Transform notation, IFT the Inverse Fourier Transform, Z(d, X) the signal X delayed of a duration d, we could write the synthesized signal as follow:

$$\frac{\sum_{i \in S} W_i Z(L_i, IFT(C_i)) + \sum_{j \in P} W_j Z(L_j, IFT(C_j)) + \sum_{i \in N} IFT(\|C_i\| FT(B W_i))}{\sum_{i \in S \cup N} W_i + \sum_{j \in P} W_j}$$

Where :

- S is the set of singular events index.
- N is the set of noise events index.
- P is the set of new periodic events index.

4.5 Residual Filtering.

An all-pole is used to filter the residual. The parameters of the all-pole filter are interpolated according to the same rules as use for the inverse filtering. At last a de-emphasis is done on the signal for respect of the spectral balance.

5. PAMAMETERS MODIFICATIONS.

To modify parameters, using an original events parameters sequence P defined by :
$P = \{ P_i = (E_i , A_i, C_i, F_i) \}$.

We construct, according to modification rules a new events parameters sequence P':
$P' = \{ P'_i = (E'_i , A'_i, C'_i, F'_i) \}$.

5.1 Rate Modification.

To perform a rate modification α we use these rules :

$P'_i = ((\alpha.L_i, T_i), A_i, C_i, F_i)$

5.2 Melodic Modification.

To perform a rate modification β we use these rules :

$P'_i = (E_i , A_i, C_i, \beta.F_i)$

5.3 Spectral Modification.

Given f a spectral anamorphosis. Using the procedure described precedently (3.4), we compute new all-pole parameters A'_i, approximing :

$$\frac{1}{A(e^{-j f(\omega)})}$$

We define C'_i by : $C'_i(\omega) = C_i(f(\omega))$

Then we have : $P'_i = (E_i , A'_i, C'_i, F_i)$

6. RESULTS.

The rules are so that singular events are not deleted or duplicated when performing any modifications. Figure 4 show the comparison of the original signal and a synthesized signal with a melody of 80% of the original.

We can analyse the synthesis of residual signal using inverse Fourier transform of $C_i(\omega)$ as a MA model of residual signal. We have then an ARMA model. We show at Fig 5 the spectral envelope resulting from AR modeling of first envelope, the envelope resulting from AR modeling of first envelope, that is the AR part of the model. We show at Fig 7 the product of $\|C_i\|(\omega)$, MA part of the model, with the AR part of the model. (sampling frequency is 16 Khz and all models are of order 20).

Informal listening of synthesis has given very good results. Without prosodic modifications, synthesis and original signals are quite indiscearnable. Prosodic modifications give great naturalness results even for women voices.

7. CONCLUSION.

Our model allows for a precise and perceptually satisfying description of speech signal (phase parameter, magnitude spectrum accurate description). Simultaneously, it provides great flexibility for prosodic modifications (pitch parameter, high level description). A further improvment of the model is the inclusion of a frequency depending voicing decision in description of pseudo-periodic events.[Rodet 87] Such a parameter would increase the range of possible realistic prosodic modifications.

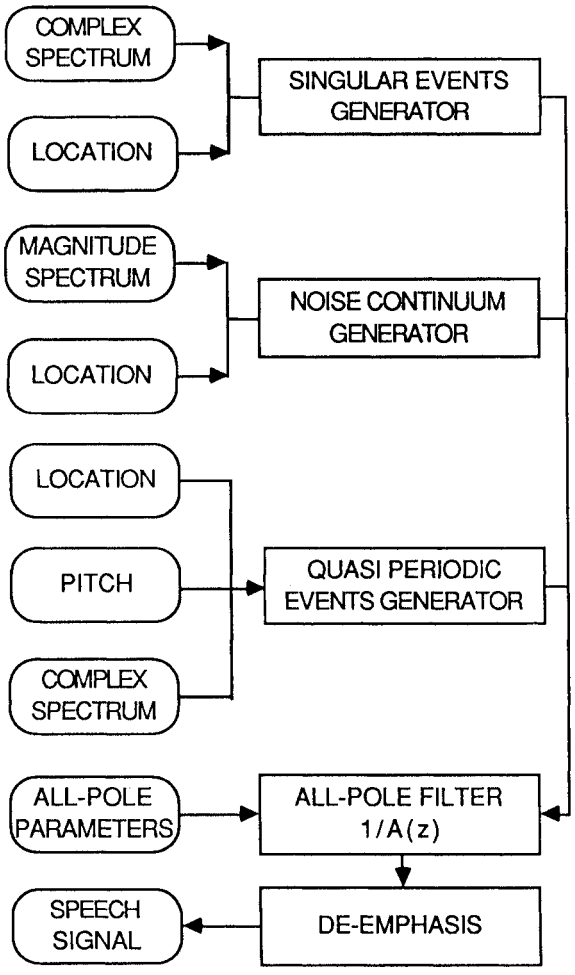

Fig 1 Production Model.

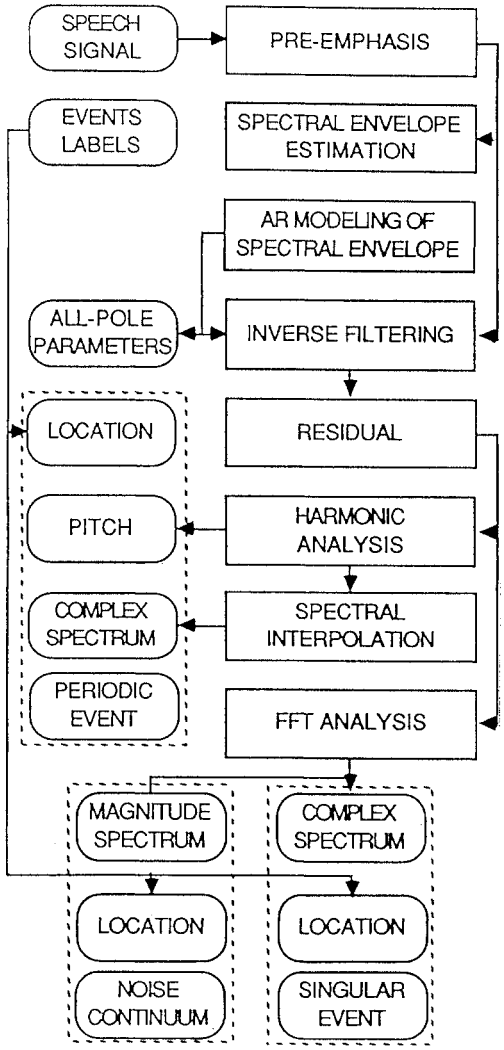

Fig 2 Analysis Process.

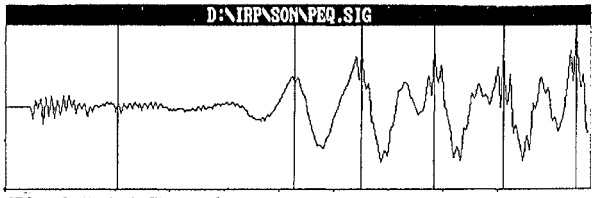

Fig.3 Labels Example.

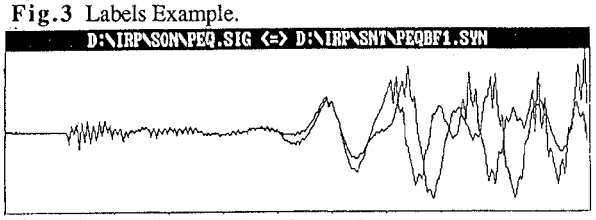

Fig.4 Synthesis vs Original.Comparison

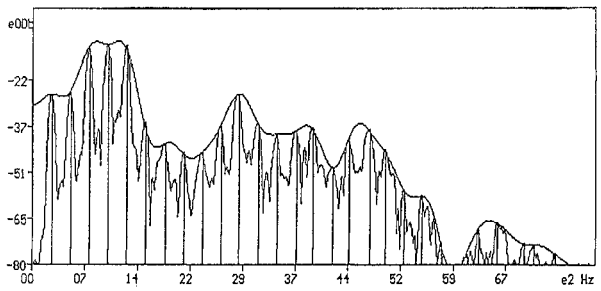

Fig 5 Discrete Cepstrum Envelope P_i (ω)

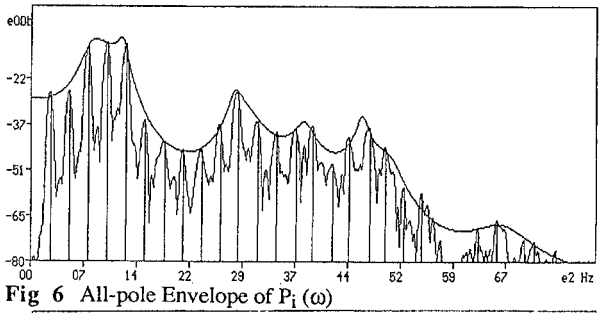

Fig 6 All-pole Envelope of P_i (ω)

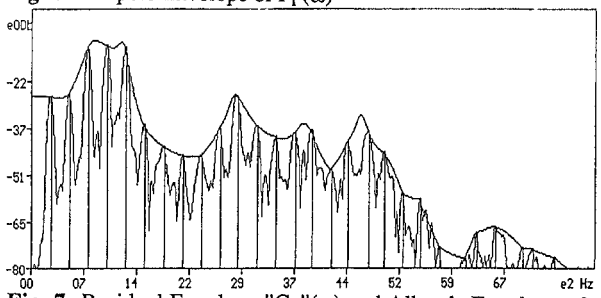

Fig 7 Residual Envelope $\|C_i\|(\omega)$ and All-pole Envelope of P_i (ω) Product.

BIBLIOGRAPHY

[Abry 84] C. Abry, L.J. Boe, R. Descout, "An events choice for a speech signal temporal organisation", 14e JEP.SFA..

[Atal 82] B.S. Atal, J. Remde. "A new model of LPC excitation for producing natural-sounding speech at low bit rates", IEEE IC ASSP, Paris,1982.

[Charpentier 89] F. Charpentier, E. Moulines, "Pitch-Synchronous Waveform Processing Techniques For Text-To-Speech Synthesis Using Diphones", Eurospeech, Paris, Sept 1989.

[Galas 90] T. Galas, X. Rodet,"An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sound signals", ICMC, Glasgow, Sept 1990.

[Itakura 68], F. Itakura, S. Saito, "Analysis Synthesis Telephony based on the Maximum Likelihood Method" Proc. 6eme Intern. Congr. Acoust., Tokyo, C17-20, 1968.

[Imai 79], S.Imai, Y. Abe, "Spectral envelope extraction by improved cepstral method", Trans. IECE, vol.J62-A, n°4, pp217-223 1979 (in japanese)

[Rodet 87], X. Rodet, P. Depalle, G. Poirot, "Speech Analysis and Synthesis Methods based on spectral envelopes and voiced/unvoiced functions", European Conf. Speech Tech., Edinburg, Sept 1987.