DIPHONE SOUND SYNTHESIS BASED ON SPECTRAL ENVELOPES AND HARMONIC/NOISE EXCITATION FUNCTIONS

X. Rodet(1)(2), Y. Depalle (1)(3) & G. Poirel(1)

ABSTRACT

A new method for sound synthesis-by-paste is described here. It is based on the use of

diphones to calculate the parameters of an original source-filter model. We first explain what
diphones synthesis is and how we implemented and improved it, then we present the

source-filter model that we control from diphones.

In any synthesis method, one has to make the distinction between calculation of samples

giving (in context) parameters and calculation of the parameters themselves, which should

be updated at a lower rate, something we call the frame rate or parameter rate (typically lower

than 200 Hz).

Considering the nature of parameters, existence of an appropriate analysis method is of

importance (Frequency Modulation is an example where analysis is not straightforward).

Thus we have worked out a very precise analysis technique described later in this article.

However, the level of interaction of parameters used in synthesis is also of importance. FPT

analysis for instance is well suited for additive synthesis but parameter values (frequency and

amplitude of harmonics) are closely dependent on the fundamental frequency of the sound.

Consequently, using this technique, one in principle cannot change the fundamental frequency

of a sound without losing some control over the timbre of the sound. On the contrary, the

parameters that we use, spectral envelopes, fundamental frequencies and harmonic/noise

characteristics are more closely related to independent perceptual features.

BIPHONE SYNTHESIS

Let us consider now the variations of parameters. A steady state sound can usually be

described with only one set, or vector, of parameters. But synthesizing transient sounds

requires a sequence of successive parameter vectors, one for each successive frame.

Calculating this vector veries time makes synthesis of transient sounds much more difficult.

Diphone synthesis overcomes this difficulty by using a trend sequence of parameter vectors

which describe the transition from one "sound" to another. Suppose we note sounds with

(1) IRCAM, 31 rue Saint-Merri, 75004 Paris, France. Tel: (1) 47 77 12 33 fax: 48-27, 48-12, usage at

adresses devanier@iraunec/rodet/1440
(2) Also with L.A.F.G.I.A., Université Paris-6, France.
(3) Also with École Supérieure d'Électricité, Evry, France.
(4) This research was partly supported by CNRS-Lannion, France.
...we note that the dipphone corresponding to the transition from U to V, VW for the transition from V to W, etc. Support as well that we consider silence as a sound named 's', and let SU designate an attack from silence to sound U. A simple note could then be described by the sequence (meaning concatenation) of three dipphones: \(SUUVW\) where SU is the dipphone which accounts for the steady portion of the note. However, this process of concatenation needs to be refined as we will see later.

Dipphone synthesis is particularly interesting because we can get dipphone data by analysing natural sounds (or eventually computing dipphone data in advance). An extremely complex transition, such as the attack of a sound, can be captured into a dipphone record, i.e. the sequence of parameter vectors produced by an analysis algorithm applied to successive frames of the transition signal. Our analysis consists of spectral envelope estimation and coding of the excitation function in terms of harmonic/period bands. An excitation function is represented by the value of the harmonic/noise index versus frequency. Spectral envelopes can be represented by coefficients \(p_n\) of a lattice filter (in LPC synthesis), or in terms of peaks of the envelope (similar to formants) given by their frequency, amplitude and bandwidth. Thus, a dipphone record is a sequence of vectors of peaks and vectors of indexes.

**OVERLAP-ADD OVER FRONTIERS**

As we mentioned earlier, raw concatenation of dipphone records is too crude. First, since dipphone records are extracted from natural sounds, acoustic characteristics coded in the last vector of a dipphone record such as UV, is often different from acoustic characteristics coded in the first vector of a VW dipphone record. This would make acoustic characteristics discontinuous at the frontier UV/WV. Second, even if this discontinuity is solved, we still need to avoid the influence of a dipphone that is not of an acoustic ganster onto its neighbors. We do this by overlap and interpolation of acoustic characteristics within a time segment that is centered on the VN frontier as explained on Figure 1. Let \(P_{U}(t)\) and \(P_{W}(t)\) be some acoustic parameter of dipphones UV and VW respectively, and let \(t_1\) and \(t_2\) be the limits of the interpolation region centered on the frontier UV at time \(t\). The result of overlap interpolation is the acoustic characteristic \(P_{over}(t)\). To compute \(P_{over}(t)\), we first calculate

\[
P_{U}(t) = P_{U}(t) + P_{W}(t)
\]

where \(r(t)\) is a weighting function increasing from 0 to 1 on the interpolation region. We then calculate \(P_{over}(t)\) as a linear segment from the point \((t_1, P_{U}(t_1))\) to the point \((t_2, P_{W}(t_2))\). Finally

\[
P_{over}(t) = r(t) P_{U}(t) + (1-r(t)) P_{W}(t)
\]

where \(r(t)\) controls the amount of mutual influence over the frontier UV.

In fact, we use other control parameters. For instance, \(s_{t_2}\) controls the speed of transition UV (such as an attack) instead of \(P_{U}(t)\), we use \(P_{over}(t)\). Consequently, the transition is faster or slower depending on whether \(s_{t_2}\) is greater or smaller than 1. These rules used to join two dipphones provide us with great variability in the description of sound evolution: transitions can exhibit more or less minor influence \(r(t)\), can be slower or faster \(s_{t_2}\), etc... independent of the duration of sound segments. Our method thus allows precise control over a large variety of timing as well as interpolation and phase effects.

ICMC Proceedings 1988
SMOOTHING PEAK TRAJECTORIES

The smoothing procedure described in the last paragraph is easily implemented on such parameters as the K\textsubscript{r} lattice-filter parameters or the associated Log Area Ratios (LAR) parameters which behave even better under interpolation. The procedure becomes much more difficult with spectral peak, the number of which is not constant since some peaks can appear or disappear during transitions. We have designed an algorithm to join two opposite UV and VW and smooth their peaks trajectories over some region. To explain it, let us consider a digraph record, i.e. a sequence of peak vectors, as a certain number of chains of mountains separated by valleys (Figure 2). The last vector of UV and the first vector of VW cannot be very different since they approximately represent the same sound V: their main chains of mountains can only slightly differ by their frequency, amplitude and bandwidth values. Thus, we can consistently match chains of mountains over the UV/VW frontier and interpolate on the corresponding values of frequency, amplitude and bandwidth. If some chains of mountains happen to have no correspondent over the frontier, it fades out (or in) smoothly.

HARMONIC/NOISE ANALYSIS

Having coded spectral envelopes of sounds, we still have to code the characteristics of the LPC residual. LPC synthesis is often heard as "buzzy", this being due, in part, to the pulse excitation which is "too periodic". Many sounds effectively exhibit a short-time spectrum which is partly harmonic (paras), partly random (noise). Consequently, we have developed an algorithm giving a harmonic/noise index in frequency bands covering the entire spectrum, that we apply to the residual signal. This algorithm is detailed in [Model 87]. The result of the analysis of one frame is a vector of indexes corresponding to frequencies from 0 up to half the sampling rate. An example of the output is shown in Figure 3 under the frequency-axis. Each harmonic is represented by a short vertical double line and noise bands are represented 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, harmonic/noise indexes lead to a remarkable improvement in the "natural" quality.

SMOOTHING HARMONIC/NOISE INDEXES

Smoothing these indexes over the frontier of two sounds is not so simple because only values 0 and 1 makes sense and because we must avoid a brusque discontinuity in the overall harmonic/noise characteristics of a frame. Thus, we partition the frequency range into a small number of large bands, each one being labeled as "totally harmonic" or "totally noise" according to the statistics of one or the other characteristic in that band. The smoothing is therefore simpler. It first consists in having these bands narrowed to a zero frequency interval or expanded towards a larger interval in order to smooth the discontinuity over the frontier (Figure 4). These indexes are calculated by interpolation in the frontier region and rounded to 0 or 1.

SPECTRAL ESTIMATION AND PEAK EXTRACTION

We use the classical source-filter approximation to model sound production. An autoregressive (AR) model of signals provides a good spectral estimation when the number of poles is high enough. In addition, resolution methods are simple and efficient. Since the signal is non-stationary, a recursive adaptive lattice method has been selected. The K\textsubscript{r} reflection coefficients are also guaranteed to be less than one in magnitude so that the synthesis filter is stable.
To accurately model fast transitions, we use a window with good properties for our analysis, i.e. one which is rather flat on the right end and smoothly damped to zero on the left end. It leads to an extremely accurate estimation, even on a segment as short as 20 ms [Roden72].

We code the spectral envelope in terms of its maxima or peaks. They are characterized by their central frequencies, amplitudes and 3-dB bandwidths. The method is detailed in [Roden87].

RECALCULATING FILTER COEFFICIENTS FROM PEAKS

Usage of a parallel formant synthesizer leads to well known problems such as zeros created between two formants. Furthermore, as we mentioned before, peaks may appear or disappear during transitions and this can hardly be handled by a parallel synthesizer. In consequence, we prefer the use of a serial filter, but calculating filter coefficients from peak values then requires more effort.

From peak values we first compute the magnitude transfer function \( M(f) \) that should be applied to the excitation function. It is the sum of the individual magnitude transfer functions corresponding to each peak. This calculation is explained in [Roden76].

Then we look for a serial filter with transfer function \( T(f) \). Let the \( n \) peaks be:

\[
( \mathbf{f}_k, A_k, B_k ) , \quad 1 \leq k \leq n
\]

and let \( f_s \) be the sampling rate. We can consider that a peak is due to a conjugate pair of poles with normalized center pulsation

\[
\omega_k = 2\pi f_s \mathbf{f}_k / \mathbf{f}_0
\]

data radius

\[
\rho_k = \exp \left( \pi \mathbf{B}_k / \mathbf{f}_0 \right)
\]

We can first determine a filter with similar peaks by computing a product of second order sections, each one with a conjugate pair of poles defined by \( \omega_k \) and \( \rho_k \):

\[
p(z) = \frac{1}{\mathbf{f}_0} \left( 1 - 2\rho_k \cos(\omega_k) z^{-1} + \rho_k^2 z^{-2} \right)
\]

which leads to:

\[
p(z) = 1 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_m z^{-m}
\]

But this filter has coefficients \( b_i, i = 1, \ldots, m \), though it has peaks of the same frequency \( f_k \) and the same bandwidth \( B_k \), has peak amplitudes different from \( A_k \) because of the serial filter hypothesis. Its magnitude transfer function

\[
P(f) = \left| \mathbf{P}(f) \right| = \left| \mathbf{P}(f) \right|
\]

is fairly different from the expected magnitude transfer function \( M(f) \). But, due to the similarity of peaks in \( P(f) \) and \( M(f) \), the ratio \( Q(f) = M(f)/P(f) \) is a smooth function of \( f \). We can then easily compute a few uncorrelation coefficients \( c_k, 0 \leq k \leq S \), associated with \( Q(f) \). Then, by solving the so called normal problem using Durbin-Levinson's method, we can get the coefficients \( d_i \) of a filter with a magnitude transfer function \( D(f) \) which is very close to \( Q(f) \).

Hence, the predefined filter \( D(f) \) has a magnitude transfer function very close to the expected magnitude transfer function \( M(f) \):

\[
D(f) \cdot P(f) \approx M(f)
\]
and the synthesis filter coefficients $g_k$ are obtained from the product:

$$G(z) = (1 + d_1 z^{-1} + \ldots + d_k z^{-k})(1 + p_1 z^{-1} + \ldots + p_k z^{-k})$$
$$G(z) = 1 + g_1 z^{-1} + \ldots + g_{2k+1} z^{-(2k+1)}$$

Ultimately, since we use a lattice filter, the $g_k$ coefficients are converted (through Levinson recursion) into the corresponding partial coefficients $k_m$ to be used in the filter.

**APPLICATIONS**

Different examples of analysis/synthesis with or without $f_0$ modification, show that a high quality sound can be obtained, essentially without the usual defects of LPC, such as 'buzzing'. We thus have a good LPC vocoder particularly suited for musical applications since its parameters (Spectral peaks, fundamental frequency, harmonic/noise index) are easy to manipulate.

We have applied diphone synthesis to musical sounds, singing voice and speech. In the case of speech, we have built a dictionary of diphones. Segmentation of natural diphones is still manual, the rest of the analysis being automatic. However, the dictionary is not complete because the number of necessary diphones range from approximately 1000 to 4000 if many diphone variants are taken into account. Each diphone record in the dictionary consist of the following files:

- peaks versus time (frequency, amplitude, bandwidth),
- micro-melody, i.e. deviation of fundamental frequency $f_0$ from mean $f_0$ in percentage,
- harmonic/noise index versus frequency and time.

Input data of the synthesis program is as follows:

```
<phoneme_label> <duration> <f0> <intensity>
```

```
<phoneme_label> <duration> <f0> <intensity>
```

**CONCLUSION**

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

The second purpose is synthesis-by-rule. Segments of sound, singing voice, or speech, such as diphones, are coded in terms of spectral peaks and harmonic/noise indexes. The results are concatenated and smoothing rules are applied to produce high quality synthesis. The input data for synthesis are the labels of successive sounds with their duration, fundamental frequency, and eventually other features of the sound quality. The method benefits from both the diphone approach and from the peak (formant) description (Roden[8]).

317

ICMC Proceedings 1988
REFERENCE


[D'Ale87] C. D'Alessandro, X. Rodet, Fonctions d'Onde Formantiques, Extraction des Paramètres et Synthèse vocale au moyen des 16èmes journées d'étude sur la parole de la Société Française d'Acoustique, Hammamet, Tunisie, October 81.


Fig. 7. Smoother peak trajectories

ICMC Proceedings: 1988
Fig. 3. Log magnitude FFT and LPC spectra of 15 ms of sound. The harmonic/noise decision is also displayed under the frequency axis.

Figure 4. Smoothing Harmonic/noise indexes.

Frequency

\[\text{time} \]

\[t_1\quad t_f\quad t_2\]

Frequency

\[\text{time} \]

\[t_1\quad t_f\quad t_2\]

ICMC Proceedings 1988