Spectral Envelope Estimation and Representation for Sound Analysis–Synthesis

Diemo Schwarz *(schwarz@ircam.fr)*, Xavier Rodet *(rod@ircam.fr)*

IRCAM – *Centre Georges Pompidou · 1, place Igor–Stravinsky · 75004 Paris · France*

Abstract

Spectral envelopes are very useful in sound analysis and synthesis because of their connection with production and perception models, and their ability to capture and to manipulate important properties of sound using easily understandable "musical" parameters. It is not easy, however, to estimate and represent them well, as several requirements must be fulfilled. We discuss the strengths and weaknesses of the estimation methods *LPC*, *cepstrum*, and *discrete cepstrum*, and evaluate the representations *filter coefficients*, *sampled*, *break-point functions*, *splines*, and *formants*. The proposed high-level approach to spectral envelope handling is followed in software developed at IRCAM, which makes some important applications of spectral envelopes in the domain of additive analysis–synthesis possible.

1 Introduction

A spectral envelope is a function giving amplitude over frequency, which is the envelope of the magnitude of a short-time spectrum (STS) [1][2]. Spectral envelopes are well suited for musical sound synthesis because of their connection with production models such as the source–filter model,[1] as well as with the perception of musical sounds, e.g. timbre. Moreover, spectral envelopes offer a simple and concise representation of important sound properties which largely simplify the control of synthesis models.

Even though the term spectral envelope is commonly used for the envelope of the magnitude of the STS only, we will also consider *generalised spectral envelopes* [1] for the phase of the STS, and for the frequency of nearly-harmonic partials as a function of their harmonic number. Several important properties of sound are thus captured in a simple and powerful representation. As sinusoidal partials and non-sinusoidal components (called the *residual noise*) in the voice and musical instruments are created by different mechanisms, their spectral envelopes have to be treated separately at all steps from estimation to synthesis.

The high-level approach to spectral envelope handling proposed here focuses on the generality of the representation of spectral envelopes, and tries to abstract from specific analysis or synthesis methods.

1.1 Properties of Spectral Envelopes

Envelope fit: A spectral envelope is a curve which envelopes the magnitude STS, i.e. it wraps tightly around it, linking the peaks of the sinusoidal partials (see discrete cepstrum in figure 1 for an example), or passing close to the maxima of non-sinusoidal spectra.

Smoothness: A certain smoothness of the curve is required: it should not oscillate irratically (fluctuate too wildly over frequency), but give a general idea of the distribution of energy of the signal over frequency.

Adaptation to fast spectrum variations: A spectral envelope is defined relative to a short segment of the signal (typically between 10 and 50 ms). When the STS varies rapidly from one analysis frame to the next, the spectral envelope should follow precisely.

2 Estimation

Estimation is the task of deriving spectral envelopes from a given signal. The requirements for an estimation method are that the properties of spectral envelopes be satisfied, plus the requirement of:

Robustness: The estimation should yield precise and smooth spectral envelopes for a wide range of signals with very different characteristics.

2.1 Methods

Linear Predictive Coding: The methods in the efficient auto-regressive class (often termed LPC) essentially build up a spectral envelope as the transfer function of an all-pole filter with *order* p poles.

Cepstrum: The cepstrum smoothes the STS of a signal by low-pass filtering its log magnitude as if it was a signal. The higher the *order*, the more accurately the envelope follows the fast variations of the STS, and hence the smoothing effect is reduced.

Discrete Cepstrum: The discrete cepstrum spectral envelope [3] is computed from distinct points in the frequency–amplitude plane. These points, not necessarily regularly spaced, are the spectral peaks of a STS,

1. As an example for the constraints imposed by the source–filter model, in speech or in the singing voice, the spectral envelope is almost independent of the pitch. However, if we transpose a vowel up by one octave, e.g. by resampling, the spectral envelope will be transposed also. This effect sounds quite unnatural since formants are shifted up one octave, which corresponds to shrinking the vocal tract to half of its length. Obviously, this is not the natural behaviour of the vocal tract. To avoid this, the spectral envelope estimated from the untransposed sound has to be restored after transposition. (See sections 4 and 5 for how to accomplish that.)

most often the sinusoidal partials found by additive analysis. The preciseness of discrete cepstrum estimation can be balanced by using a *logarithmic frequency scale* above a given break frequency, similar to the *mel scale*. This reflects the frequency resolution of the human ear, which is coarser for high frequencies.

2.2 Comparison

Figure 1 shows the weaknesses of LPC and cepstrum estimation: Both descend down into the space between the partials, when they are spaced far apart, as for high-pitched sounds. For low-order estimation, the LPC is too smooth, and misses some of the peaks. The cepstrum has the problem of averaging the spectrum, i.e. it does not link the peaks either.

All these problems are avoided by the discrete cepstrum method. Nevertheless, LPC and cepstrum are still very well applicable to the residual noise, where the discrete cepstrum cannot be used.[2]

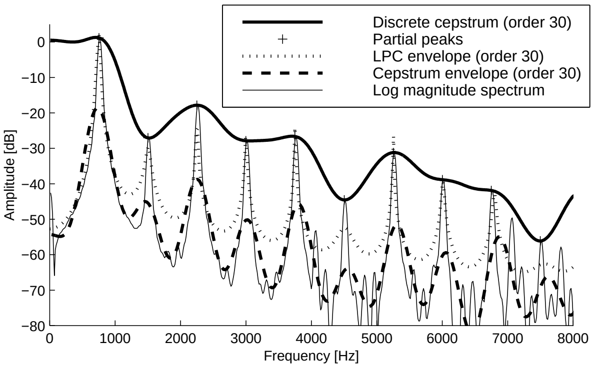

Figure 1: Comparison of LPC, cepstrum, and discrete cepstrum spectral envelope estimation.

3 Representation

As we have seen above, the various estimation methods result in very different parameterizations of spectral envelopes. However, their unified, high-level representation is essential for their use in musical synthesis and for the flexibility of transformations and further processing. After giving the requirements for representation, we present several representations and their comparison.

3.1 Requirements

Preciseness: The representation has to describe an arbitrary spectral envelope (obtained by estimation, or given manually) as precisely as possible.

Stability: The requirement of stability mandates that the representation be resilient to small changes in the data to be represented. Small changes, e.g. in the presence of noise, must not lead to large changes in the representation, but must result in equally small changes. Stability is of great importance considering that the data to be represented can result from various different estimation methods, or from manual input, and that some noise is always present.

Locality in frequency: This requirement states that it be possible to achieve a local change of the spectral envelope (not affecting the amplitude further away from the point of manipulation) by a simple change in the representation parameters.

Flexibility and ease of manipulation: The representation must be flexible enough to allow various manipulations. It must be easy to specify manipulations with an exactly defined desired outcome, e.g. a certain formant location that has to be reached in voice synthesis. For the manipulative abilities to be really useful for musical applications, the relationship between the parameters of the manipulation and the effect on the spectrum has to be easily understandable.

Speed of synthesis: The representation should be usable for sound synthesis as directly as possible, without first having to be converted to a different form at high computational costs. This requirement is heavily dependent on the type of synthesis.

Space in memory: The representation must not take up too much memory, which is important for file storage and even more so for transmission.

Manual input: The representation should be easy to specify manually, e.g. by drawing a curve or placing primitive shapes, or by textual input of parameters.

3.2 Proposed Representations

Filter coefficients: We can directly use the parameters from estimation, be it the cepstral or one of the several types of LPC coefficients.

Sampled representation: The continuous spectral envelope from estimation (or given directly) is sampled at n frequency points, which can be equidistant or logarithmically spaced.[3]

Geometric representations: These try to describe the curve of the spectral envelope with fewer points than the sampled representation not spaced at equidistant frequencies. They can be given as a piece-wise linear, or **break-point function** (BPF), which represents a function by linear segments linking the break points $P_i(\mathit{frequency}_i, \mathit{amplitude}_i)$, or as **splines**, which provide for quadratic or cubic interpolation of each section between the given points. In the latter case, the points are placed on the maxima, minima, and inflection points of the spectral envelope.

Formants: In the voice, the maxima of the spectral envelope convey the main part of the perceptual information concerning the vocal tract and the timbre. These peaks result from resonances in the vocal tract (or any acoustic resonator) which are called formants. Since several formants are needed to represent a spectral envelope, it has to be defined how they combine to form the envelope: by addition or multiplication. These two com-

2. Robustness of estimation of speech spectral envelopes can be improved using a composite envelope which is discrete-cepstrum-estimated from the voiced part below the *maximum voiced frequency* (the frequency of the highest sinusoidal partial), and LPC-estimated from the unvoiced part above, as described in [4] and implemented in [5].

3. Equal care as when sampling audio signals has to be taken to assure that n is high enough to avoid aliasing of the rapidly varying components of the continuous spectral envelope.

binations correspond, respectively, to the parallel and serial structure of synthesis filters. Their different properties have been largely discussed, e.g. in [6]. There are three convenient ways to represent formants:

1. **Formant waveforms** (FOFs, from *Forme d'Onde Formantique*) represent a formant as an elementary waveform [7]. Several FOFs add up to build the desired spectrum (typically 5–7 for a voice). The frequency-domain parameters are center frequency, amplitude, bandwidth, and skirt width (which can be controlled independently from the bandwidth); the time-domain parameters are phase, excitation time, and attenuation time. Although FOFs are a very precise way to define a spectrum for singing voice and music synthesis (in the CHANT system [8]), they bear more information than is needed for the representation of a spectral envelope.

2. **Basic formants** are a simpler way to describe formant spectral envelopes, using the parameters center frequency c_k, bandwidth b_k, and amplitude a_k (in dB). With these parameters, the spectral envelope of the basic formant k can be defined as:

$$v_k(f) = \frac{10^{\frac{a_k}{20}}}{1 + \left(10^{\frac{3}{20}} - 1\right)\left(\frac{c_k - f}{b_k/2}\right)^2} \tag{1}$$

It approximates very well the magnitude transfer function of a two-pole filter, which is the usual model of a resonance. The final spectral envelope is the sum of the basic formants $v_k(f)$.

3. Finally, precisely representing a real-life spectral envelope as formants is often difficult. However, the approximate locations of formants are fairly well known. This motivates defining **fuzzy formants** as regions within a sampled spectral envelope where we assume that a formant exists.[4] A fuzzy formant is specified by the lower bound, the upper bound, and the center corresponding to the frequency of the formant peak. Additionally, we identify formants with a label, such that they can be associated into formant tracks.

3.3 Comparison

The table shows a comparison of the representations with a score ($++, +, \mathsf{O}, -, --$) indicating fulfillment of the requirements from section 3.1. (The preciseness requirement is not listed, as it is fulfilled by all methods).

Representation	Stability	Locality	Flexibility/ Ease of Manipulation	Speed of Synthesis TD/FD	Space	Manual Inp.
Filter coef.[5]	$++$	$-$[6]	$--$ / $-$[6]	$++$ / O	$+$	$--$[6]
Sampled	$++$	$++$	$++$ / $+$[7]	$-$ / $++$	O	$+$[7]
Geometric[8]	$-$[9]	$+$	$+$ / $++$[10]	$-$ / $+$	$+$	$++$
Formants[5]	$-$[9]	$+$	$++$ / $++$	$+$ / O	$++$	$++$[11]

4 Synthesis

In synthesis from scratch, a spectral envelope is given directly as part of the synthesis parameters. In resynthesis, an input signal is modified so as to respect the desired spectral envelope.

To apply a spectral envelope, we can use **filtering**, where the spectral envelope has to be converted to filter coefficients for time-domain filtering, or to a transfer function for filtering in the frequency-domain[12]. Various types of filters are abundantly described in the literature, e.g. in [9].

In **additive synthesis**, the synthetic signal is a sum of sinusoidal partials with amplitudes according to the sinusoidal spectral envelope, and of a residual noise the spectral density of which is given by the noise spectral envelope. The residual can be synthesized by filtering white gaussian noise. For the sinusoidal part we have to replace or crossfade the amplitude of a partial with the value of the spectral envelope at its frequency.

The **FFT^{-1} method of additive synthesis** [10] avoids the computational cost of the classic oscillator method. It uses the inverse Fourier transform of a STS, allowing a speed gain of 10 to 30. It is implemented in various musical sound synthesis systems [10][11][12]. Synthesizing the residual is easily and inexpensively done while constructing the STS before transformation: just add random values in the desired frequency bins.

5 Applications

A function library in *C* and various programs have been developed at IRCAM, which allow spectral envelope estimation and their application to sound transformation and synthesis [2]. Using the proposed high-level approach to spectral envelopes, we can simplify the problem of **controlling sinusoidal partials for additive synthesis**, and manipulating them in a sensible way. This has often been addressed by specifying the change of every single parameter over time by break-point functions (e.g. in [13]). Since the number of partials can easily rise into the hundreds, modifications are tedious. Moreover, doing valid manipulations in regard of signal processing and from a musical perspective is not obvious, and the parameters are interdependent. In [14] it is

4. This knowledge can come from manually labeled source material (a recording of the voice with annotations of the phonemes that are uttered), or from automatic formant estimation.

5. There are envelopes which are not easily representable, for example the ideal low-pass filter.

6. Changing one coefficient changes the envelope at all frequencies.

7. They are not that easy to manipulate, because their high locality demands that all the new values at all the frequencies be given.

8. Geometric representations don't model the spectral envelope in a way relevant to its properties in relation its source signal, but simply as a curve in euclidian space. More specifically, interdependencies between the given points, that arise from the signal character of the spectral envelope are not taken into account automatically.

9. Small changes in the envelope can cause a sudden change of the maxima/formants found. (However, with fuzzy formants, such an instability is not damaging.)

10. There is a tradeoff between ease of manipulation and preciseness: When there is a point that governs a large area that can thus be manipulated easily, the preciseness can suffer because a large portion of the curve will be changed.

11. For specifying spectral envelopes manually, e.g. for the precise synthesis of the voice, formants are well suited.

12. FD-filtering is done e.g. with IRCAM's phase vocoder *SuperVP*.

suggested to use spectral envelopes to control the amplitudes of the partials for resynthesis. This drastically reduces the number of parameters, provides us with parameter sets which are easily understandable (e.g. formants), and renders frequency and amplitude control independent from each other.

Modeling the residual noise part by filtering white noise with spectral envelopes renders this component of sound accessible to manipulation. This has not been possible in the sampled signal representation of the residual. The most significant advantage, however, lies in the unified high-level handling of noise and harmonic parts: because the spectral envelope of the residual noise is represented in the same way as that of the sinusoidal part, a manipulation can affect both parts synchronously, if this is desired. Sinusoidal and noise spectral envelopes are used in IRCAM's real-time synthesis system *jMax* [12] using FFT^{-1}.

To perform **modification and synthesis of the singing voice** in a sensible manner, the constraints posed by the speech organs have to be taken into account (e.g. when transposing). Also, many aspects of the expressivity of the singing voice depend on the spectral envelope, i.e. on timbral variations like spectral tilt, rather than on pitch and loudness alone. With the methods of morphing between spectral envelope and formants described in [15], a new type of high quality synthesis of the voice is possible: To preserve the rapid changes in transients (e.g. plosives), and the noise in fricatives, these are best synthesised with the harmonic sinusoids + noise model, controlled by spectral envelopes in sampled representation. For precise formant locations in the steady part of vowels, the formant representation is used. With morphing between fuzzy and precise formants, it is then possible to interface the excellent generation of vowels by FOF synthesis, as used in the CHANT synthesiser [8], with the flexibility of general additive synthesis, for instance in the generalized graphical synthesis control program DIPHONE [16].

6 Conclusion

In the context of computer music, the control of spectral envelopes offers the possibility to influence the timbre of a sound to a great degree, allowing composers to obtain a desired effect or characteristic of a sound by the use of flexible, unconstrained, high-level representations. To the performer, the real-time application of spectral envelope manipulation greatly enhances expressivity through easily understandable and "musical" parameters, i.e. parameters that pertain to a model.

The previous sections lead to the observation that each representation has its strong points. To keep maximum flexibility, we have to use all of them,[13] and combine them in an object-oriented class hierarchy.

With the software developed at IRCAM, sophisticated new sound transformation and synthesis methods, also in real-time, are possible. Note that all the programs use the standardized, open, and extensible *Sound Description Interchange Format* (SDIF, cf. [17][18]) to facilitate the exchange of spectral envelope data with well-defined semantics [2] between programs, hardware architectures, and institutions. With more and more analysis–synthesis tools being ported to SDIF, this will create important synergetic effects in research and creation.

For an in-depth discussion of spectral envelopes, see also the forthcoming book [15].

References

[1] X. Rodet, Ph. Depalle, G. Poirot. Speech Analysis and Synthesis Methods Based on Spectral Envelopes and Voiced/Unvoiced Functions. *European Conf. on Speech Tech.*, 1987.

[2] D. Schwarz. *Spectral Envelopes in Sound Analysis and Synthesis*. Diplomarbeit Nr. 1622, Universität Stuttgart, Fakultät Informatik, Stuttgart, Germany, 1998.

[3] Th. Galas, X. Rodet. Generalized Functional Approximation for Source–Filter System Modeling. *Proc. Eurospeech*, 1991.

[4] Y. Stylianou, J. Laroche, E. Moulines. High Quality Speech Modification based on a Harmonic+Noise Model. *Proc. EUROSPEECH*, 1995.

[5] M. Campedel Oudot. *Étude du modèle sinusoïdes et bruit pour le traitement de la parole. Estimation robuste de l'enveloppe spectrale*. Thèse, ENST, Paris, 1998.

[6] J. N. Holmes. Formant synthesizers: Cascade or Parallel. *Speech Communication*, vol. 2, 1983.

[7] X. Rodet. Time-Domain Formant-Wave-Function Synthesis. *Computer Music Journal*, Fall 1984.

[8] X. Rodet, Y. Potard, J.-B. Barrière. The CHANT–Project: From the Synthesis of the Singing Voice to Synthesis in General. *Computer Music Journal*, Fall 1984.

[9] R. W. Hamming. *Digital Filters*. Signal Processing Series. Prentice–Hall, 1977.

[10] A. Freed, X. Rodet, Ph. Depalle. Performance, Synthesis and Control of Additive Synthesis on a Desktop Computer Using FFT^{-1}. *Proc. ICMC*, 1993.

[11] X. Serra, J. Bonada, P. Herrera, R. Loureiro. Integrating Complementary Spectral Models in the Design of a Musical Synthesizer. *Proc. ICMC*, 1997.

[12] F. Déchelle, M. DeCecco, E. Maggi, N. Schnell. *jMax* Recent Developments. *Proc. ICMC*, 1999.

[13] K. Fitz, L. Haken, B. Holloway. Lemur – A Tool for Timbre Manipulation. *Proc. ICMC*, 1995.

[14] A. Freed, X. Rodet, Ph. Depalle. Synthesis and Control of Hundreds of Sinusoidal Partials on a Desktop Computer without Custom Hardware. *ICSPAT*, 1992.

[15] X. Rodet, D. Schwarz. *Spectral Envelopes and Additive+Residual Analysis-Synthesis*. In J. Beauchamp, ed. *The Sound of Music*. Springer, N.Y., to be published.

[16] X. Rodet, A. Lefèvre. The Diphone Program: New Features, new Synthesis Methods and Experience of Musical Use. *Proc. ICMC*, 1997.

[17] M. Wright et al. New Applications of the Sound Description Interchange Format. *Proc. ICMC*, 1998.

[18] M. Wright, A. Chaudhary, A. Freed, S. Khoury, D. Wessel. Audio Applications of the Sound Description Interchange Format Standard. *AES 107th convention preprint*, 1999.

13. Between most of the representations conversion is easy.