Chapter III
Audio Watermarking Through Parametric Synthesis Models

Yi-Wen Liu
Boys Town National Research Hospital, USA

ABSTRACT

This chapter promotes the use of parametric synthesis models in digital audio watermarking. It argues that, because human auditory perception is not a linear process, the optimal hiding of binary data in digital audio signals should consider parametric transforms that are generally nonlinear. To support this argument, an audio watermarking algorithm based on aligning frequencies of spectral peaks to grid points is presented as a case study; its robustness is evaluated and benefits are discussed. Toward the end, research directions are suggested, including watermark-aided sound source segregation, cocktail watermarking, and counter-measure against arithmetic collusive attacks.

BACKGROUND

Frequency-domain masking is often regarded as the standard perceptual model in audio watermarking. Below the masking threshold, a spread-spectrum watermark (e.g., Kirovski & Malvar, 2003; Swanson, Zhu, & Tewfik, 1998) distributes its energy; the same threshold also sets a limit to the step size of quantization in informed watermarking (e.g., Chou, Ramchandran, & Ortega, 2001). However, ways to manipulate sounds can go far beyond masking to deceive the human auditory system. This has been explored by a group of attackers; in 2001, participants invited by Secure Digital Music Initiative (SDMI) successfully defeated audio watermarking schemes that were then state of the art (Wu, Craver, Felten, & Liu, 2001). Typically, successful attacks exploited human insensitivity to small changes in phase, pitch, and time. Since then, many have wondered how these aspects may also be used for improved watermarking. A plethora of audio watermarking schemes thus came into existence.

Without pursuing details of all the schemes and their applications, this section aims to provide a unifying view. First, the concept of parametric
sound synthesis is introduced. Then, general design principles for watermarking with synthesis models will be stated. Finally, existing watermarking methods will be categorized according to the signal models.

**Parametric Modeling in Digital Audio: A Brief Tutorial**

Simply put, “parametric modeling” means to represent a signal comprised of a large number of samples with fewer variables called parameters. Whenever such a model can be constructed for a signal, it immediately leads to significant data compression. Historically, data compression through parametric modeling has worked very well in speech through linear prediction (Markel & Gray, 1976). Human vocalization has been studied extensively, and speech signals are nowadays routinely encoded in parameters derived from estimated motion of the vocal tract and vibration rates of the vocal cord (Shroeder & Atal, 1985).

Nevertheless, speech is not the only type of audio signals that can be parameterized, and benefits of parameterization go beyond saving the data storage space. In mid 1980s, sinusoidal modeling was independently proposed by two groups; McAulay and Quatieri (1986) sought to parameterize speech by tracking spectral peaks, and Smith and Serra (1987) devised a similar scheme to generate musically expressive effects such as stretching a signal in time without dropping its pitch.

Predictive modeling and sinusoidal modeling are not mutually exclusive; the residual component in sinusoidal modeling can be parameterized by linear prediction (Serra & Smith, 1990). The hybrid system is referred to as “deterministic plus stochastic,” because the residual component lacks tonal quality and thus is modeled as filtered noise. The deterministic plus stochastic model was refined by Levine (1998) by further decomposing the stochastic component into a quasi-stationary “noise” part and a rapidly changing “transient” part. Thus, Levine’s model led to a versatile audio coding scheme that is both efficient and expressive. Development in sinusoidal modeling has culminated in its adoption by MPEG-4 as an extension to the audio coding standard (Purnhagen & Meine, 2000). F-QIM, the main watermarking algorithm to be described in later sections, is heavily based on Smith, Serra, and Levine’s work.

In a futuristic fashion, parametric signal modeling has been used in network-based musical interaction, and languages and protocols are still being designed (e.g., Vercoe, Gardner, & Scheirer, 1998; Wessel & Wright, 2004). Often, sounds are encoded by physically meaningful parameters and synthesized through wave equations in object-oriented manners (Cook & Scavone, 1999; Scavone & Cook, 2005). As new synthesis models and music formats emerge, their impact on the audio watermarking is yet to be seen.

**Watermarking Through Parametric Modeling**

For the convenient of discussion, hereafter, a parametric model is defined as a mapping from a parameter space to a Euclidean signal space \( R^n \) (or \( C^n \)). The mapping is also referred to as a “synthesis.” Conversely, a recorded signal \( s \) can be decomposed into a perfectly parameterized part \( s \) and a residual \( r \) orthogonal to \( s \). Demonstrated in Figure 1, the triangle-shaped region \( \Xi(\Theta) \) on the left represents the image of the parameter space \( \Theta \) in the signal space \( R^n \). For an arbitrary signal \( s \) in \( R^n \), its closest neighbor \( s_{\Theta} \) is searched in \( \Xi(\Theta) \), and an inverse mapping takes \( s_{\Theta} \) back to its parametric representation \( \Theta \). The inverse mapping is also referred to as an “analysis,” or “parameter estimation.” Mathematical notations used in this chapter are listed in Table 5.

The diagram in Figure 2 shows a generic watermarking scheme based on parametric analysis and synthesis. To embed a binary watermark \( W \), the cover signal \( s \) is first decomposed into \( s + r \)