TAPESTREA: Sound Scene Modeling By Example

Ananya Misra, Perry R. Cook, and Ge Wang
Princeton University*

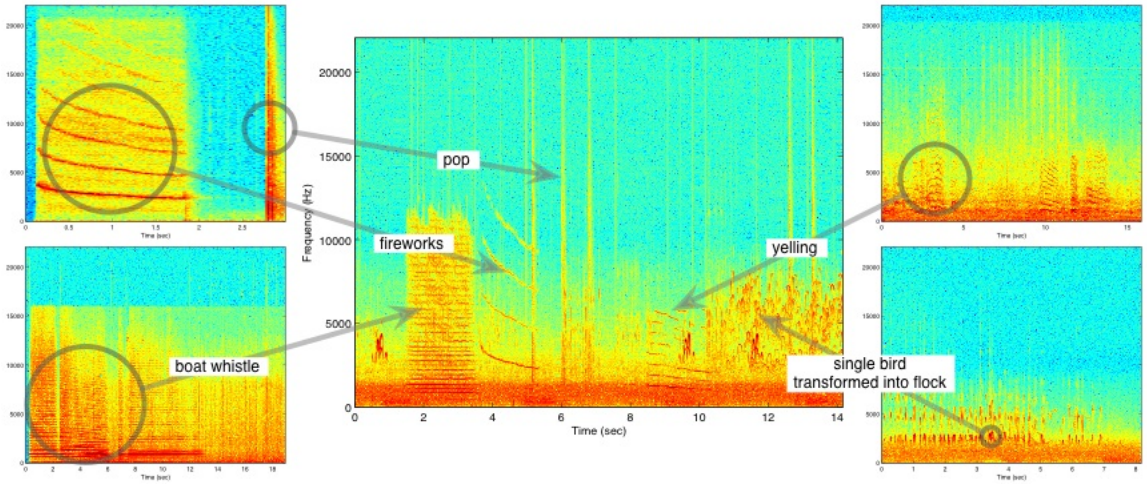

Figure 1: *A sound scene composed of background and foreground elements from several existing scenes*

1 Introduction

We present a new paradigm and framework for creating high-quality "sound scenes" from a set of recordings. A sound scene is a combination of background and foreground sounds that together evoke the sense of being in a specific environment. The ability to craft and control sound scenes is important in entertainment (movies, TV, games), virtual/augmented reality, art projects (live performances, installations) and other multimedia applications.

Existing audio production tools require "untainted" versions of sound components and frequently involve tedious event-by-event editing. No system, to our knowledge, provides an arena for truly flexible "sound scene modeling by example," where a sound scene can be composed from selected, extracted, and separated components of different existing scenes. We introduce a parametric, unified framework of analysis, transformation and synthesis techniques that allow users to interactively select components from existing sounds, transform these independently, and controllably recombine them to create new sound scenes in real-time. We call this system TAPESTREA: Techniques and Paradigms for Expressive Synthesis, Transformation and Rendering of Environmental Audio.

2 Techniques and Paradigms

Our approach follows from the notion that sound scenes are composed of discrete events as well as background sound, and these are best modeled separately. A sound scene is separated into the following components: (1) *deterministic events*: composed of sinusoids, often perceived as pitched events, such as a bird chirp or a voiced vowel, (2) *transient events*: brief non-sinusoidal events, such as footsteps, (3) *stochastic background*: the "din" remaining after the removal of foreground events, such as wind or street noise.

Our parametric analysis interfaces let a user extract selected instances of any component type from a given sound scene, and save them as *templates* for future use in corresponding transformation and synthesis interfaces.

Our system benefits from employing separate analysis and synthesis algorithms for each component type. We apply spectral modeling [Serra 1989] to extract deterministic events, and resynthesize them with optionally massive time and frequency transformations. Transient events are isolated by locating sudden increases in time-domain signal energy, and are replayed directly or with time-frequency transformations. Both event types are removed from the original sound to obtain the background. Deterministic events are removed by sinusoidal track detection, while removed transient events are "filled in" or replaced using wavelet tree learning [Dubnov et al. 2002] of nearby transient-free segments. An improved wavelet tree learning technique also synthesizes continuous, non-repeating stochastic background sound, similar to the extracted background template.

A user, having extracted templates, can parametrically resynthesize them with specific individual real-time transformations. The system also offers structures for explicitly placing events in time at many granularities, and for synthesizing repeating events at controllable periodicity, density, and random-transformation ranges, useful for generating crowd sounds from a single template. Together, these enable the production of a wide range of sound scenes, of any desired length, from a static and limited set of recordings.

3 Contributions

Our main contributions include: (1) techniques and paradigms for interactive template selection and extraction, (2) techniques for parametrically transforming components independently, (3) a framework for flexible resynthesis to create novel sound scenes, (4) interfaces to facilitate each task in the analysis and synthesis pipeline. Most significant are the new approach, system, and interface for modeling sound scenes by example.

References

DUBNOV, S., BAR-JOSEPH, Z., EL-YANIV, R., LISCHINSKI, D., AND WERMAN, M. 2002. Synthesizing sound textures through wavelet tree learning. *IEEE Computer Graphics and Applications 22*, 4.

SERRA, X. 1989. *A System for Sound Analysis / Transformation / Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University.

*e-mail: {amisra, prc, gewang}@cs.princeton.edu