論文の内容の要旨

論文題目　Exploiting regularities in natural acoustical scenes for monaural audio signal estimation, decomposition, restoration and modification
（音環境に内在する規則性に基づくモノラル音響信号の推定・分解・復元・加工に関する研究）

氏名　Jonathan Le Roux
ルルー　ジョナトン

A crucial problem for many audio engineering applications is that most, if not all, real world situations they face are adverse ones, with strong non-stationary background noises, concurrent sources, brief interruptions due to glitches or missing packets, etc. Humans however are able to achieve a great robustness in their perception and understanding of the complex acoustical world that surrounds them, relying on statistical regularities in the original sources and the incoming stimuli.

The goal of this thesis is to propose a statistical approach to the analysis of such natural acoustical scenes, based on models of the regularities of the acoustical environment. Our main strategy is to systematically focus on a general mathematical formulation of the problem based on an objective function, so that the various subtasks can be effectively solved as well-posed constrained optimization problems, and to allow for easy extensibility of our work into other signal processing algorithms involving a statistical framework. Such a statistical approach involves solving mainly three subproblems: inference of what is happening in an acoustical scene as the best explanation of the distorted, mixed, and incomplete observations given models of the environment; reconstruction of incomplete observations based on these models; acquisition of these models from the data. We tackle all of these problems, following a common procedure: design of appropriate models and constraints; formulation of the task as an optimization problem; derivation of an effective optimization method.

After reviewing related works in Chapter 2, we start our work by introducing in Chapter 3 a statistical model for voiced speech signals in the time-frequency power domain called Harmonic-Temporal Clustering (HTC). The time-frequency domain formulation enables us to explicitly make use of grouping principles inspired from humans' auditory organization to

derive a completely parametric model of voiced speech signals as constrained Gaussian mixture models with a smoothly evolving F0 contour. We also introduce a broadband noise model, based on Gaussian mixture models as well, to deal with noisy environments. We explain how to formulate scene analysis tasks as the fitting of a mixture of such models to the observed spectrogram, and derive an effective method to estimate the optimal parameters, based on the EM algorithm. We show in Chapter 4 through experimental evaluation that our method outperforms state-of-the-art algorithms in classical scene analysis tasks such as F0 estimation in clean, noisy, or concurrent environments, denoising, and source separation.

In Chapter 5, we explain how scene analysis based on statistical models can be extended to deal with incomplete stimuli through an auxiliary function method. Meanwhile, we study the theoretical relation of this auxiliary function method with the EM algorithm in the particular case of Bregman divergences. We show through experimental evaluation that the proposed method enables to simultaneously perform the analysis of an underlying acoustical scene such as a polyphonic music signal and to reconstruct its missing part.

We then note that although we may gain by discarding the difficult to model phase part when working in the time-frequency magnitude domain, we also lose in several ways. First, if resynthesis is necessary, the absence of phase information needs to be dealt with by estimating the phase from the available information, i.e., the magnitude spectrogram. The estimation of a phase which corresponds well to the magnitude spectrogram is crucial to avoid very disturbing perceptual artifacts in the resynthesized signal. Second, we lose the additivity of signals, as cross-terms in the square of a sum are in general not equal to zero. Third, phase may actually be, for some classes of sounds, a relevant cue which is worth being exploited. In all cases, working in either the complex time-frequency domain or the time domain is a natural answer to deal with the problem. We present two frameworks to do so.

The first one, which we present in Chapter 6, is based on a careful study of the particular structure of complex STFT spectrograms. Due to the redundancy of the STFT representation, an arbitrary set of complex numbers in the complex time-frequency domain is not guaranteed to be what we call a "consistent" spectrogram, i.e., the STFT spectrogram of an actual time-domain signal. We derive a mathematical characterization of consistent spectrograms, and from it a cost function to measure the consistency of an arbitrary set of complex numbers in the complex time-frequency domain. We use this cost function to build an algorithm for phase reconstruction from magnitude spectrograms, and show that it is both more flexible and more efficient than the state-of-the-art method. Moreover, we note that the cost function we derive is a natural candidate to define a prior distribution on complex spectrograms, and as such likely to be used in a wide range of signal processing algorithms in the future.

The second framework, shift-invariant semi-NMF (SSNMF), is presented in Chapter 7. It is based on a direct modeling of the signal waveform, in the time domain, simply assuming that the observed waveform is the superposition of a limited number of elementary waveforms, added with variables latencies and variable but positive amplitudes. The model is more general than the HTC model presented in earlier chapters, in the sense that it is less

constrained: it can represent any kind of sound, and is not limited to harmonic ones. A sparseness prior is used on the amplitudes to ensure that the elementary waveforms capture meaningful information recurring at various time instants. We derive an optimization algorithm for this model, and show that it can be used to effectively recover recurring templates together with their activation times from the waveform of a mixture, even in the difficult case where the various templates overlap, with examples in audio signals and extracellular recordings.

Finally, we investigate the unsupervised acquisition of models based on the data, observing that, although much can be obtained using tailored models based on prior knowledge, what we can get from them will be limited by the quality and appropriateness of that prior knowledge. We first explain how the SSNMF framework presented in Chapter 7 performs a sort of data-driven learning, and how it could be used to learn compact basis for natural sound signals. Then, noting that an often overlooked but important issue when performing time-frequency analysis is to determine the analysis parameters, we consider in Chapter 8 the unsupervised learning of time-frequency analysis filterbanks. Motivated by the central role which seems to be played by modulation in auditory perception, we design a mathematical framework to investigate the hypothesis that the human ear and brain, and in particular the peripheral system, adapted for modulation analysis through a data-driven learning process. Optimizing a filterbank on speech data under a modulation energy criterion, we show that the optimized filterbank is close to classical ones, and the hypothesis pertinent.