Audio Source Separation via the Grouping of Partials in the Sum-of-Sinusoids Model

Nicholas Esterer – M.A. in Music Technology

The sparse nature of musical sounds (hereafter referred to as *signals*) allows for their succinct representation as a mixture of time-varying partials (not necessarily harmonically related). In the signal processing literature this is called the *sum-of-sinusoids* or *additive model* and the partials are called *sinusoidal components (SC)* [1]. This model is convenient for its relative parsimony while keeping common signal transformations pertinent to computer music easy to perform. Operations like time-stretching, transposition, noise-removal and filtering are simple and efficient to perform on the sinusoid model. In theory, the act of separating multiple superimposed sounds—audio source separation—should also be simple: the desired SC are chosen and the rest discarded. The difficult problem lies in choosing the group of SC that come from the same source. To this end we propose a system that extracts (numerical) descriptions of the SC and at the same time groups them into sources, based on their descriptions.

Theoretical Background and Methodology

In recent years techniques estimating the parameters of more descriptive sinusoidal models incorporate instantaneous frequency- and amplitude-modulations (FM and AM) [2]. For tractability, these parameters are only computed at a few unique times within a signal and it is assumed the resulting data points represent either noise or underlying SC belonging to a particular source. The data indicating to which source a data point belongs are missing, while we do have some information on the nature of the underlying component (the computed parameters). *Expectation Maximization (EM)* is a technique applicable in such circumstances [11]: It computes the most highly expected values of the missing data given the available information. It remains to connect the data points together in a plausible way, hopefully resulting in the original component. This can be phrased as an optimization problem: We want to find the optimal set of

connections between data points, subject to some constraints (e.g., each data point can only belong to one component, etc.). This type of problem is readily solved as a *linear program (LP)*, a technique finding renewed interest due to promising proofs of its worst-case computational complexity [3]. It has been shown that common FM [4] and AM [5] were plausible data with which to group sinusoids common to one source [6], and it is with these data that we will define a measure of optimality.

Objectives

We want to assess the applicability of the above techniques to the source separation problem. Using a set of AM and FM parameters estimated at one time, we will use the Gaussian mixture clustering algorithm (an incarnation of EM) to group parameters into sources [7]. It remains to group parameters across time. While various methods have been proposed for this (e.g., [8], [9]), motivated by its desirable properties outlined above, we will try a new method based on a technique for tracking multiple objects using LP [10]. The result are data indicating the parameter-source classification at one time, and the parameter connections across time, i.e., which parameters belong to which source. From these parameters, the original sources can be synthesized.

Contribution

This thesis demonstrates the mastering of a variety of musical signal processing techniques and their integration into a larger framework in an attempt to solve the difficult problem of audio source separation on a single channel. Previous work on this problem either required the SC to be harmonically related [12], or that their AM and FM be small (e.g., [13]). Neither are the case in our proposed method. While documenting the knowledge gained throughout the Master's program, the contribution is furthermore applicable to the domains of audio editing (e.g., unwanted sources can be discarded, effects can be applied to a single source, etc.) and analysis (e.g., measurements can be made for only particular sources).

Bibliography

[1] Serra, Xavier, and Julius Smith. "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition." *Computer Music Journal* 14, no. 4 (1990): 12-24.

[2] Hamilton, Brian, Philippe Depalle, and Sylvain Marchand. "Theoretical and practical comparisons of the reassignment method and the derivative method for the estimation of the frequency slope." In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pp. 345-348. IEEE, 2009.

[3] Boyd, Stephen, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] McAdams, Stephen. "Segregation of concurrent sounds. I: Effects of frequency modulation coherence." *The Journal of the Acoustical Society of America* 86, no. 6 (1989): 2148-2159.

[5] Li, Yipeng, John Woodruff, and DeLiang Wang. "Monaural musical sound separation based on pitch and common amplitude modulation." *Audio, Speech, and Language Processing, IEEE Transactions on* 17, no. 7 (2009): 1361-1371.

[6] Creager, Elliot. "Musical source separation by coherent frequency modulation cues." 2016. Masters Thesis, McGill University. Accessed April 18, 2016. http://digitool.library.mcgill.ca/thesisfile139201.pdf.

[7] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.

[8] McAulay, Robert J., and Thomas F. Quatieri. "Speech analysis/synthesis based on a sinusoidal representation." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 34, no. 4 (1986): 744-754.

[9] Depalle, Ph, Guillermo Garcia, and Xavier Rodet. "Tracking of partials for additive sound synthesis using hidden Markov models." In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 1, pp. 225-228. IEEE, 1993.

[10] Jiang, Hao, Sidney Fels, and James J. Little. "A linear programming approach for multiple object tracking." In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1-8. IEEE, 2007.

[11] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the royal statistical society. Series B (methodological)* (1977): 1-38.

[12] Wang, Avery Li-Chun. "Instantaneous and frequency-warped signal processing techniques for auditory source separation." PhD diss., Stanford University, 1994.

[13] Vincent, Emmanuel, Nancy Bertin, and Roland Badeau. "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription." In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 109-112. IEEE, 2008.