Interactive User-Feedback for Sound Source Separation

Nicholas J. Bryan*
Center for Computer Research in Music and Acoustics
Stanford University, CA, USA
njb@ccrma.stanford.edu

Gautham J. Mysore
Adobe Research
San Francisco, CA, USA
gmysore@adobe.com

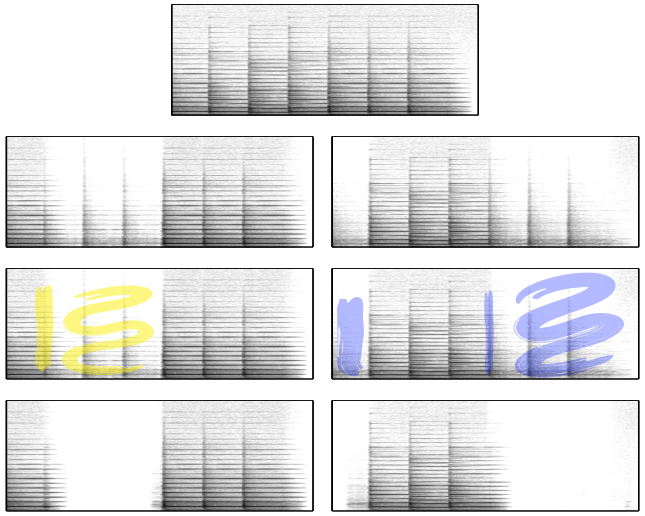

Figure 1: (First Row) Mixture spectrogram of *Mary Had A Little Lamb*. (Second Row) Initial poorly separated E notes (left) and remaining notes (right). (Third Row) Annotations overlaid indicating incorrectly separated regions. (Bottom) Refined output after user-feedback.

IUI'13, March 19–22, 2012, Santa Monica, California, USA.
*This work was performed while interning at Adobe Research.

Abstract
Machine learning techniques used for single-channel sound source separation currently offer no mechanism for user-feedback to improve upon poor results and typically require isolated training data to perform separation. To overcome these issues, we present work that applies interactive machine learning principles to incorporate continual user-feedback into the source separation process. In particular, we allow end-users to annotate errors found in source separation estimates by painting on time-frequency displays of sound. We then employ a posterior regularization technique to make use of the annotations to obtain refined source separation estimates and repeat the process until satisfied. An initial prototype shows that the proposed method can significantly improve separation quality compared to previous work and facilitate separation without isolated training data.

Author Keywords
Sound, audio, source separation, user-feedback, interactive machine learning

ACM Classification Keywords
H.5.2 [User Interface]: User-centered design, interaction styles; H.5.5 [Sound and music computing]: Methodologies and techniques, signal analysis, synthesis, and processing; I.5.4 [Applications]: Signal processing

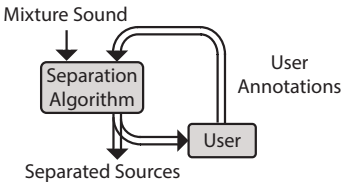

Figure 2: Block diagram of the proposed interactive source separation method.

Introduction

For many music- and audio-related tasks, it is desirable to decompose a single-channel recording of a mixture of sounds (e.g. pop song) into its respective sources (e.g. drums, guitar, vocal, etc.). Over the past decade, non-negative matrix factorization and related latent variable models have become common approaches for this purpose. While these methods can achieve high-quality separation, typically the results are less than ideal. Additionally, such methods offer no mechanism to improve upon poor results and are ineffective when no training data is available. To mitigate these issues, we first propose an interaction method to incorporate user-feedback into the separation process, and then extend a popular source separation technique to incorporate the feedback.

Interaction

To incorporate user-feedback into the separation process, we allow an end-user to initially separate a recording, listen to the separated outputs, paint on spectrogram displays[1] of the output estimates, and re-run the process until satisfied, as shown Fig. 1 as a sequence of spectrograms and Fig. 2 in block-diagram form.

When painting on the display of a particular output sound, a user is asked to identify regions that are incorrectly separated. Opacity is used as a measure of strength and color is used to identify source. For simplicity, we focus on separating one sound from another, although the method can be used to separate more than two sources at once.

Methodology

To initially separate a given recording, we use probabilistic latent component analysis (PLCA) [2, 3]. PLCA is a time-varying mixture model that decomposes audio spectrogram data into a weighted combination of prototypical frequency components over time. The frequency components and weights for each source in a mixture are estimated using an expectation-maximization algorithm. These results are then used to estimate the proportion of each source in the mixture and subsequently reconstruct each source independently.

To incorporate the user-feedback described above, the painting annotations serve as penalty weights which are used to constrain our probabilistic model via the framework of posterior regularization (PR) [1]. PR allows for efficient time-frequency constraints that would be very difficult to achieve using prior-based regularization.

To test the proposed method, a prototype user-interface was developed and used to separate several real-world sound examples. Using standard evaluation metrics, we demonstrate that the method can achieve high-quality results, and even perform well with no training data (see `https://ccrma.stanford.edu/~njb/research/iss/` for audio and video demonstrations). Such results show great promise for the use of interactive user-feedback for sound source separation.

[1] A spectrogram is an auditory display of sound which roughly depicts energy as a function of time (x-axis) and frequency (y-axis).

References

[1] Ganchev, K., Graça, J., Gillenwater, J., and Taskar, B. Posterior Regularization for Structured Latent Variable Models. *J. of Machine Learning Research 11* (2010).

[2] Raj, B., and Smaragdis, P. Latent variable decomposition of spectrograms for single channel speaker separation. In *Proc. IEEE WASPAA* (2005).

[3] Smaragdis, P., Raj, B., and Shashanka, M. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Proc. ICA* (2007).