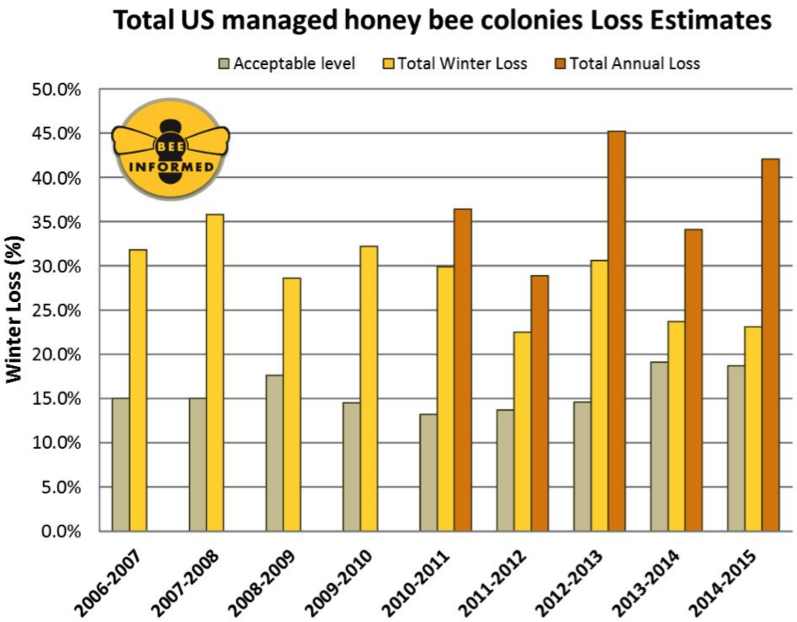
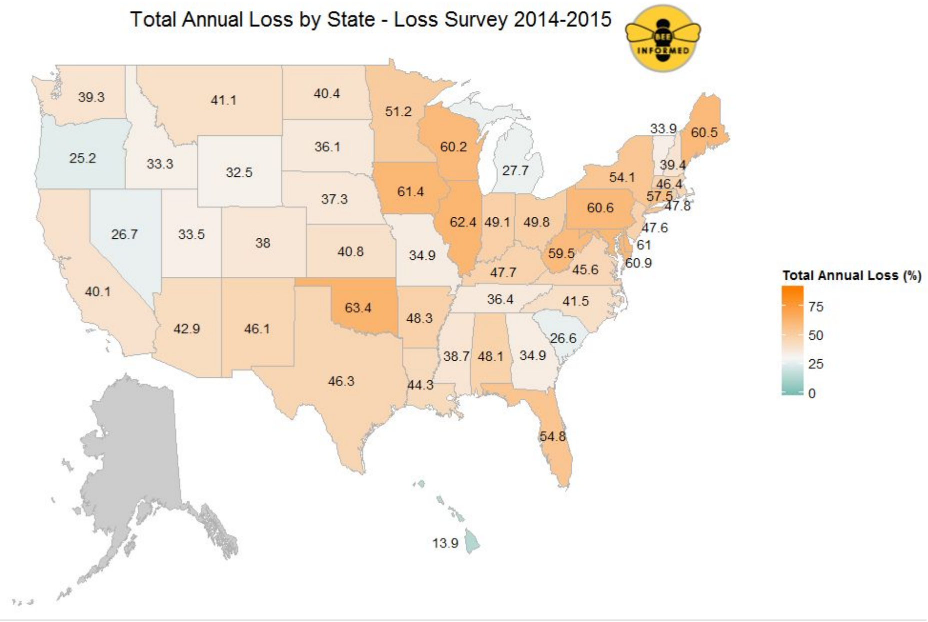
Single Channel Source Separation Applied to Beehive Audio

An Undergraduate Thesis Defense By
Dakota Murray

Colony Collapse Disorder

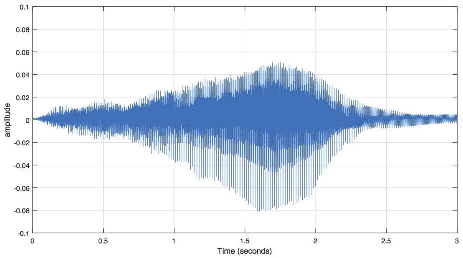


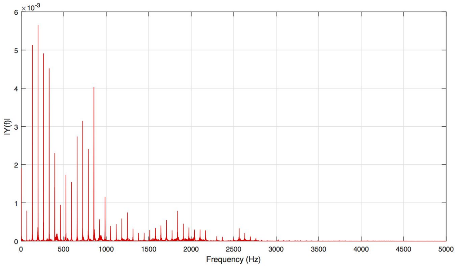


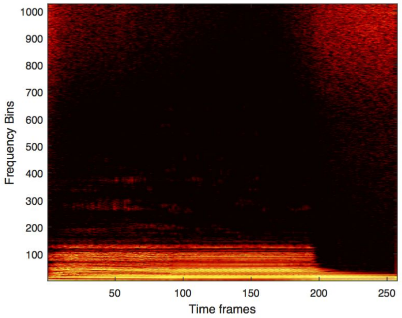
Representations of Audio

Time Series | Frequency Domain | Spectrogram

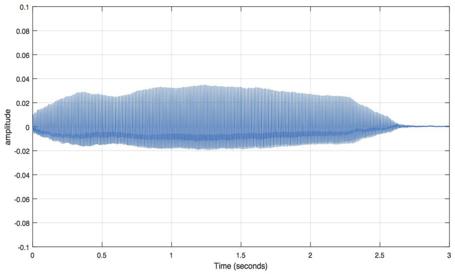
Cello

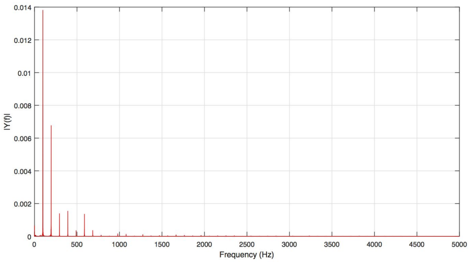


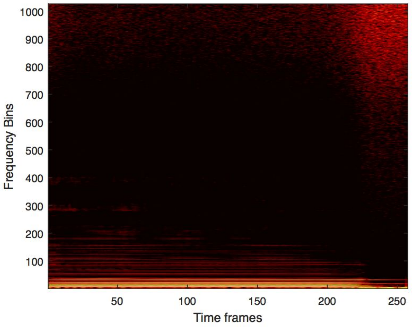




Clarinet

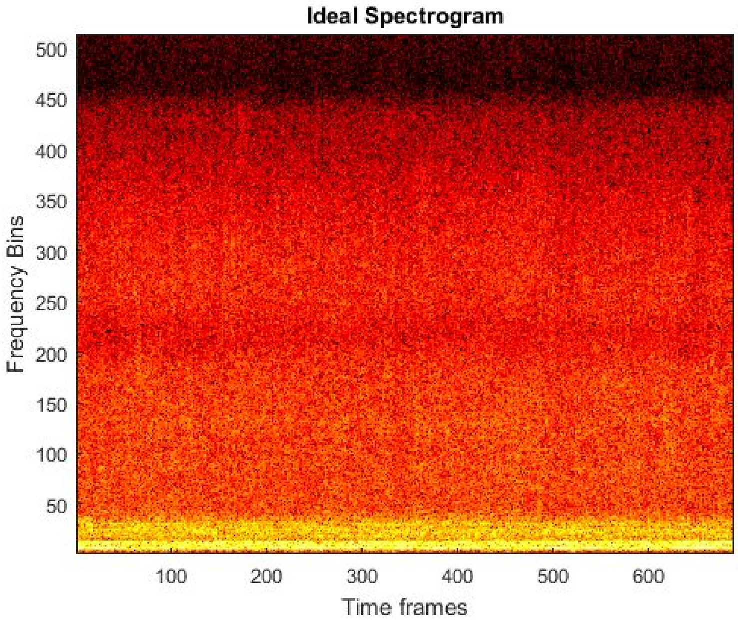






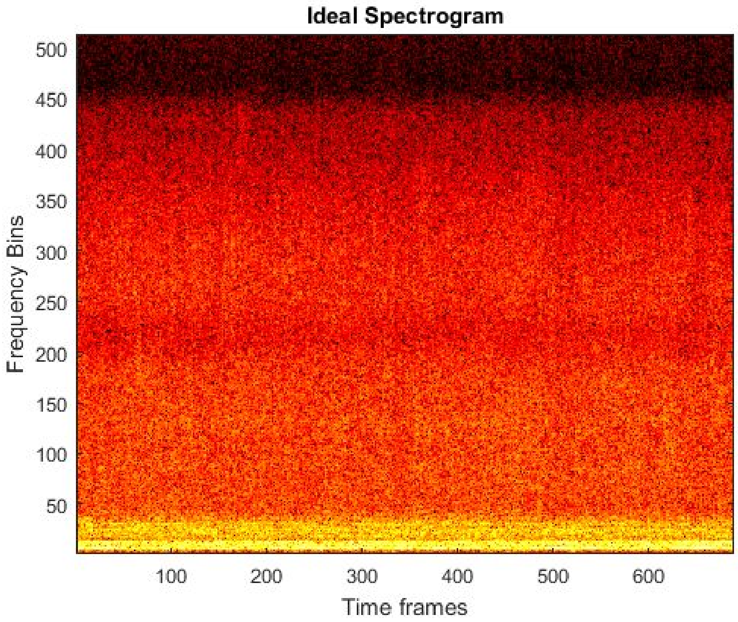
The Problem With Audio Data

What We Want

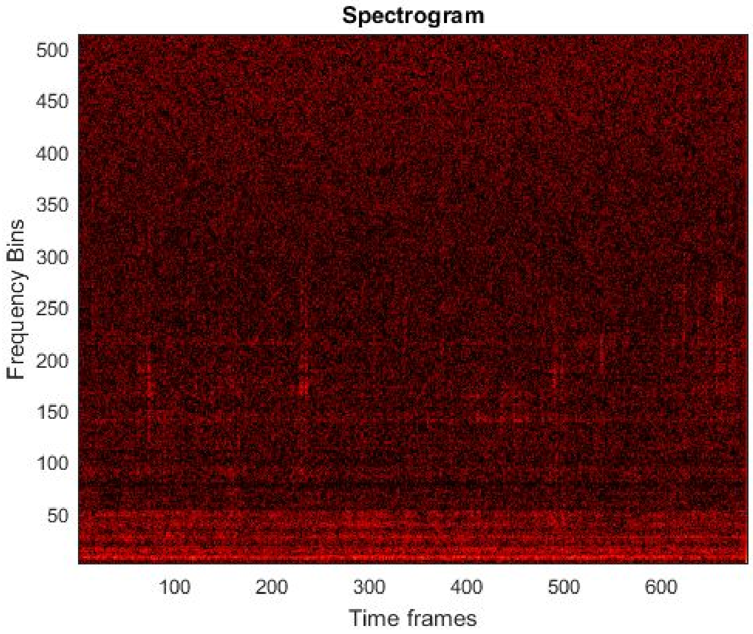


The Problem With Audio Data

What We Want

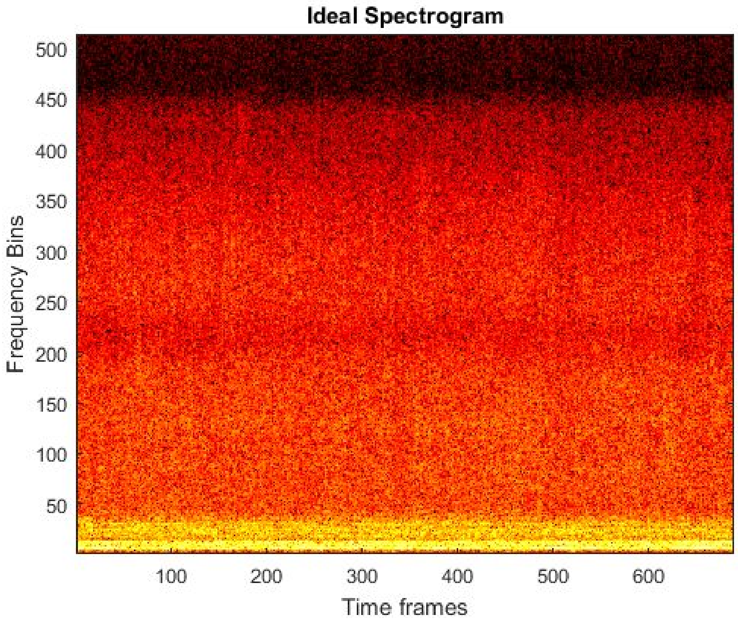


What We Have

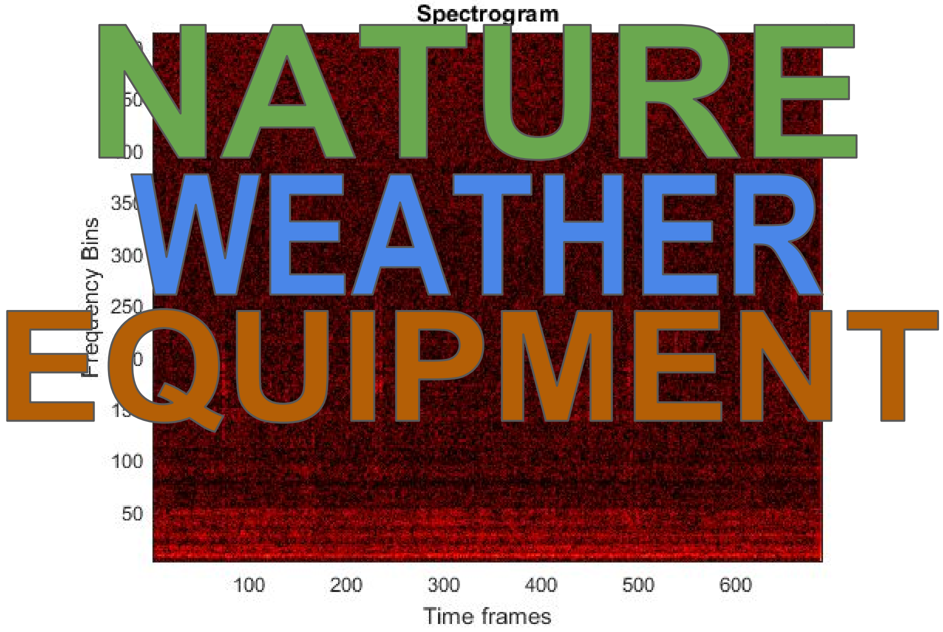


The Problem With Audio Data

What We Want

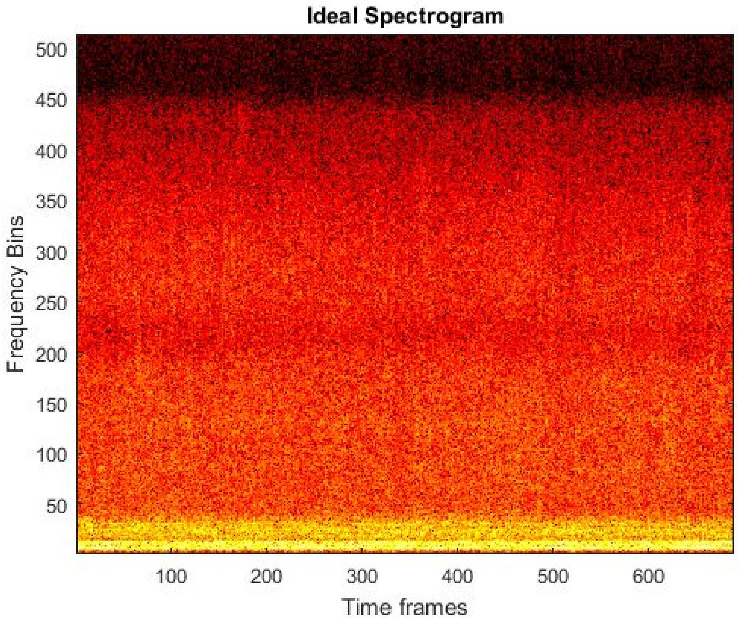


What We Have

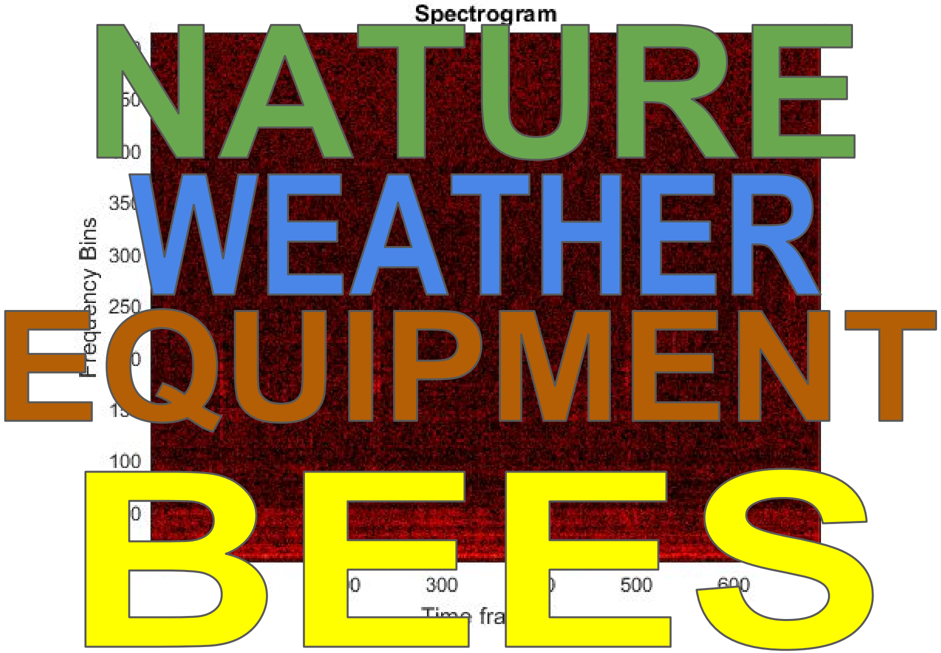


The Problem With Audio Data

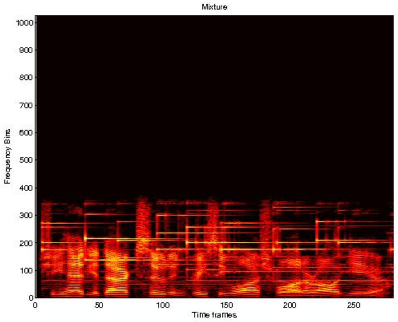
What We Want



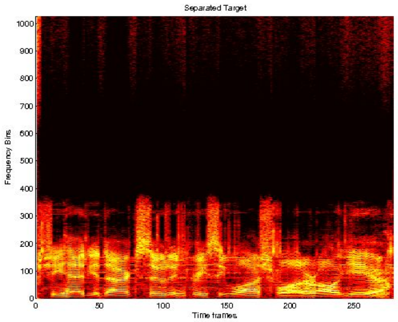
What We Have



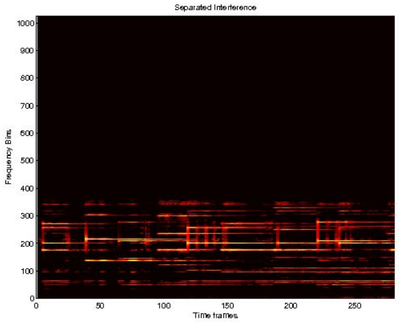
Proposed Solution: Separate the Signals



Mixture Signal



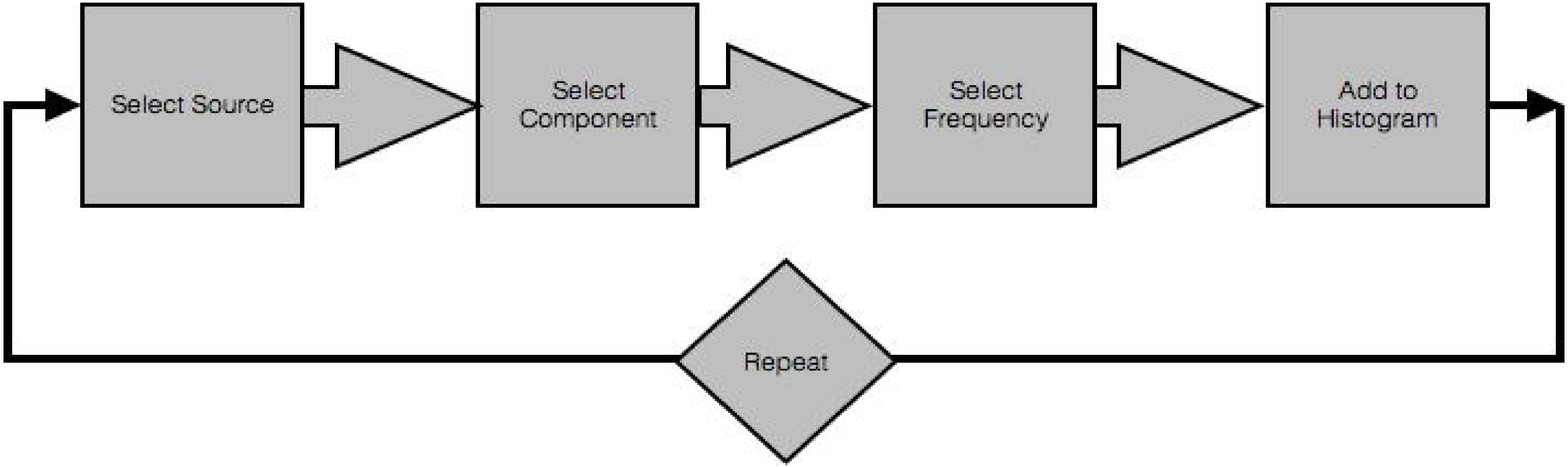
Separated Signals



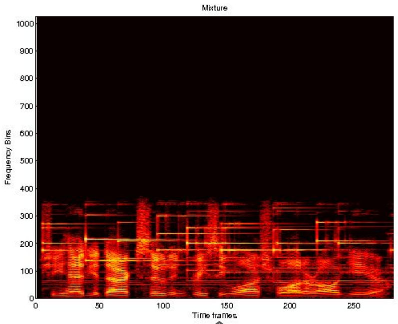
Target

Interference

Latent-Variable Model



Latent: (of a quality or state) existing but not yet developed or manifest; hidden; concealed.



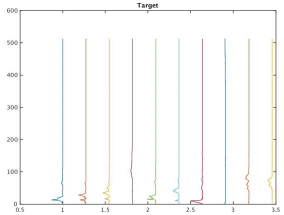
Latent-Variable Model Continued

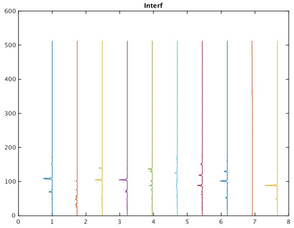
Select Source

Target (Voice)

Interference (Chimes)

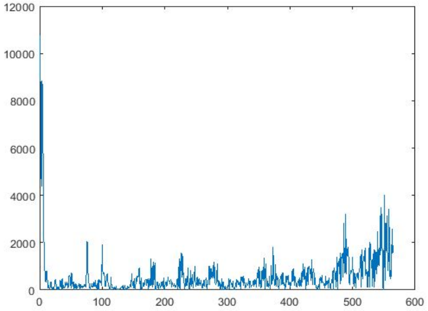
Select Component



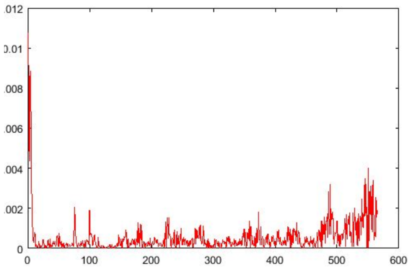


Select Frequency, Add to histogram

Repeat



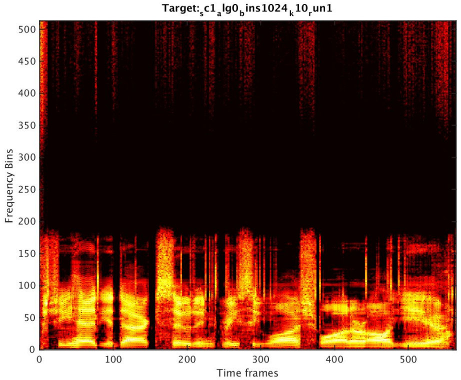
Histogram after 1 Million iterations



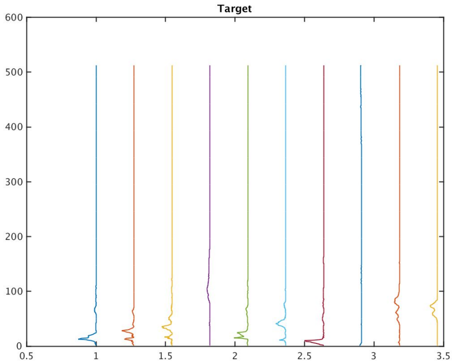
For Every Time Frame

Training Phase: Learn Components of Each Source From Training Data

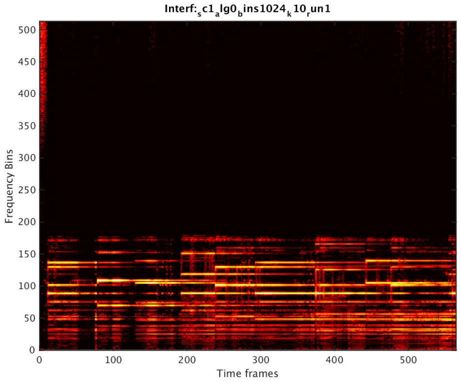
Target



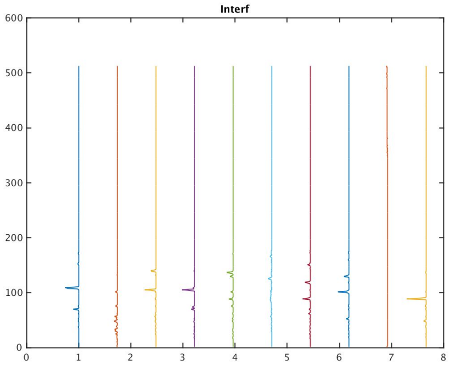
Learn



Interference

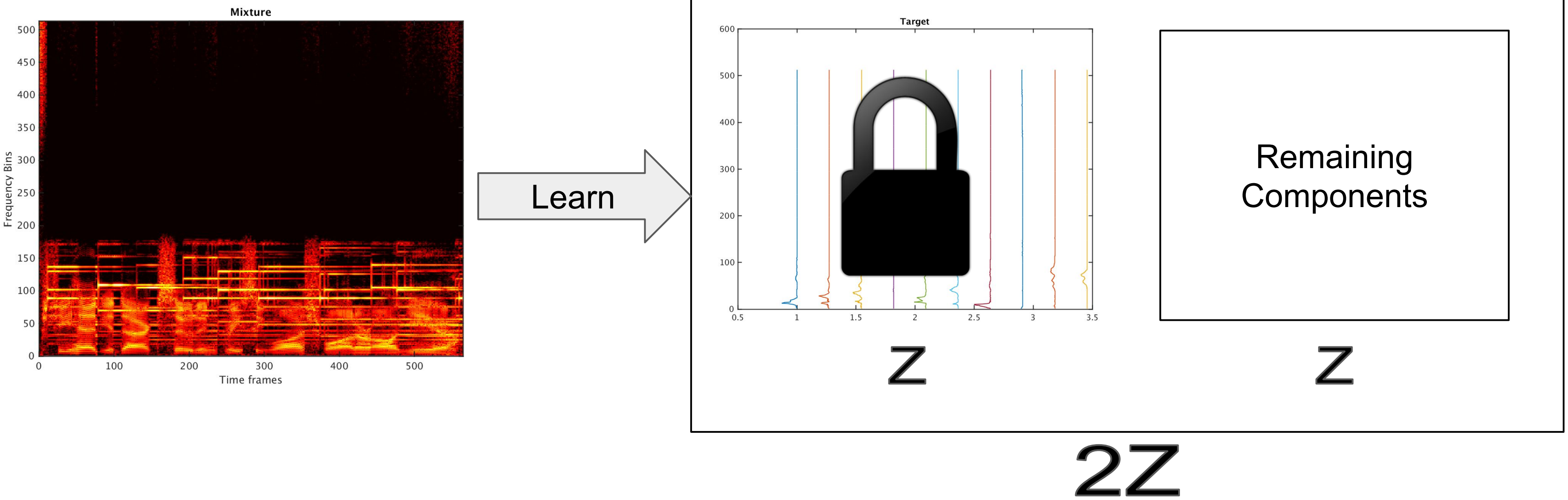


Learn



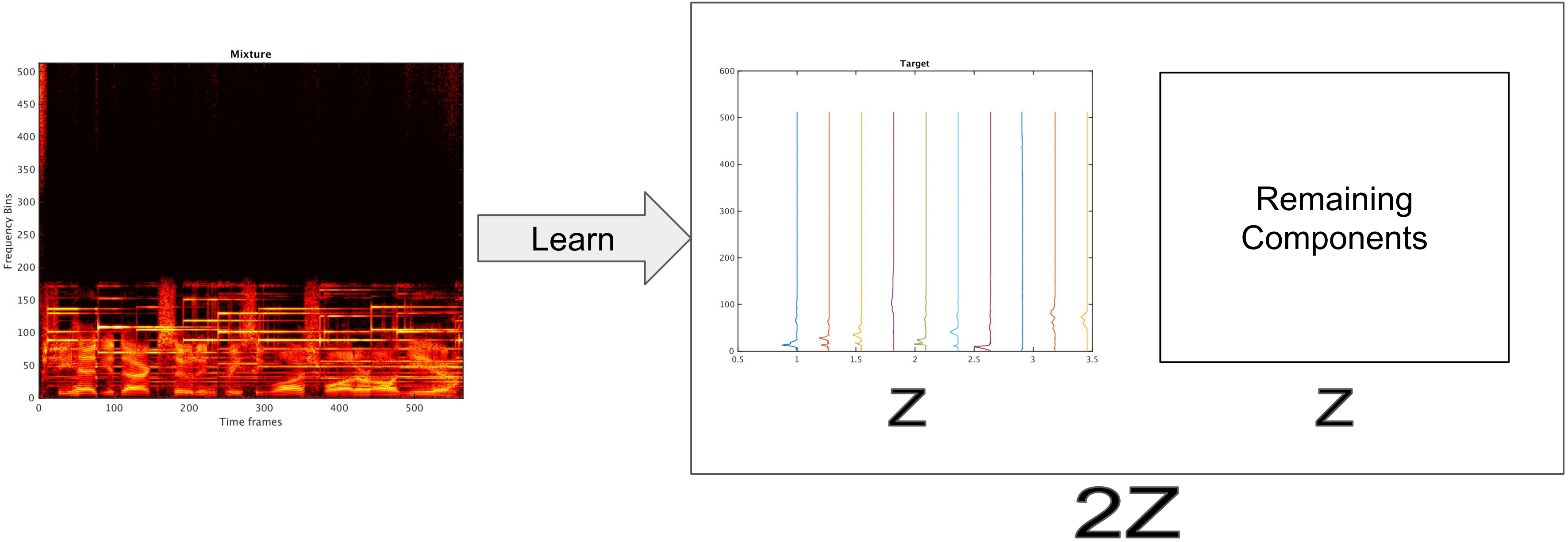
Semi-Supervised Training: Fixed Target

Learn 2Z components. First Z start as already trained Target model. Use second Z as interference model



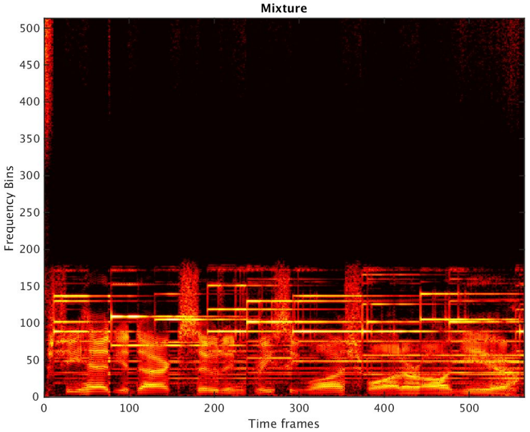
Semi-Supervised Training: Unfixed Target

Learn 2Z components. First Z start as already trained Target model. Use second Z as interference model

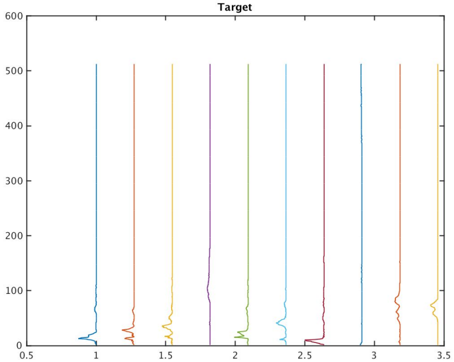


Separation Phase: Learn Contribution of Each Component to Mixture

Mixture

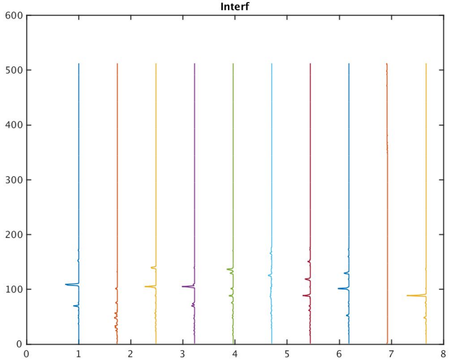


Contribution
?



Target

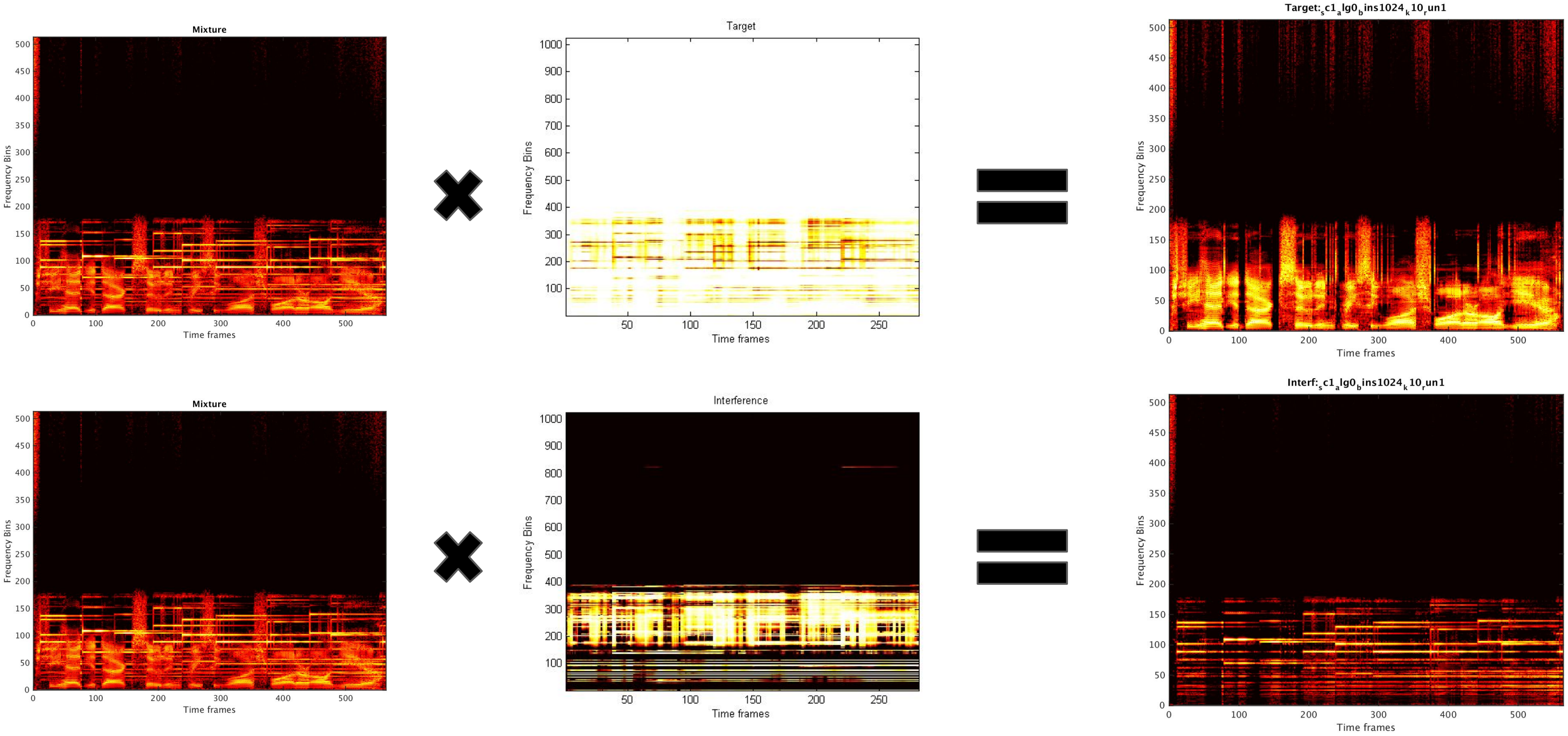
Contribution
?



Interference

Reconstruction: Calculate Weights for Source’s Contribution to Mixture

Mixture | Weights | Separated Source



Equations: Implemented in Matlab

Training

$$P_t(z|f) = \frac{P_t(z)P(f|z)}{\sum_{z'} P_t(z')P(f|z')}$$

$$P(f|z) = \frac{\sum_t P_t(z|f)N_{t,f}}{\sum_t \sum_f P_t(z|f')N_{t,f}}$$

$$P_t(z) = \frac{\sum_f P_t(z|f)N_{t,f}}{\sum_{z'} \sum_f P_t(z'|f)N_{t,f}}$$

Interference

$$P_t(s) = \frac{\sum_{z' \in \{z_s\}} \sum_f P_t(s,z|f)N_{t,f}}{\sum_{s'} \sum_{z' \in \{z_s\}} \sum_f P_t(s',z|f)N_{t,f}}$$

$$P_t(z|s) = \frac{\sum_f t(s,z|f)N_{t,f}}{\sum_{z' \in \{z_s\}} \sum_f P_t(s,z'|f)N_{t,f}}$$

$$P_t(s,z|f) = \frac{P_t(s)P_t(z|s)P_s(f|z)}{\sum_s P_t(s) \sum_{z \in \{z_s\}} P_s(f|z)P_t(z|s)}$$

Reconstruction

$$\overline{N}_{t,f}(s) = \frac{P_t(s) \sum_{z \in \{z_s\}} P_s(f|z)P_t(z|s)}{\sum_s P_t(s) \sum_{z \in \{z_s\}} P_s(f|z)P_t(z|s)} N_{t,f}$$

Research Questions

1. Does my implementation of the algorithm work at all?
2. Is the technique effective on beehive audio?
3. What parameters are most important for improving performance?
4. What parameters are most important for reducing computation time?

Scenarios

Five scenarios selected. Scenarios 3-5 contain real-world audio recorded by internal mounted beehive monitoring systems. All samples are 4 seconds long.

Scenario	Target	Interference	Source
1	Man’s Voice	Windchimes	Dr. Smaragdis’ Website
2	Isolated Beehive	Isolated Birdsong	Ideal Example, Online Audio Database
3	Real-World Beehive	Flyby	Hand-Selected, Real-World Audio
4	Real-World Beehive	Rain Striking Hive	Hand-Selected, Real-World Audio
5	Real-World Beehive	Electronic Static	Hand-Selected, Real-World Audio

Parameters: 5 Iterations Each

Variant	Supervised	Fixed Target	Unfixed Target		Semi-supervised variants should adapt to the mixture
NFFT	1024	2048	4096	8192	Higher NFFT means Higher Frequency Resolution. Lower Time Resolution
#Components	5	10	15	20	Larger number of components means a more complex model for each source

Evaluation

- **PEASS Toolkit**
  - Scores (out of 100) that correlate to perceptual quality of separation
  - Compares separation output vs originals

- Duration of each phase of the algorithm (in seconds)

OPS	Overall Perceptual Score	Holistic quality of separation
TPS	Target Perceptual Score	Quality of separated target
IPS	Interference Perceptual Score	Quality of separated interference
APS	Artifact Perceptual Score	Score of artifacts introduced to separated signals by the algorithm

Training Duration	Duration of training phase (in seconds)
Separation Duration	Duration of separation phase (in seconds)

Results

Answering the Research Questions

1. Does my implementation of the algorithm work at all?
2. How effective is the technique on beehive audio?
3. What parameters are most important for improving performance?
4. What parameters are most important for reducing computation time?

Does the Implementation Work?

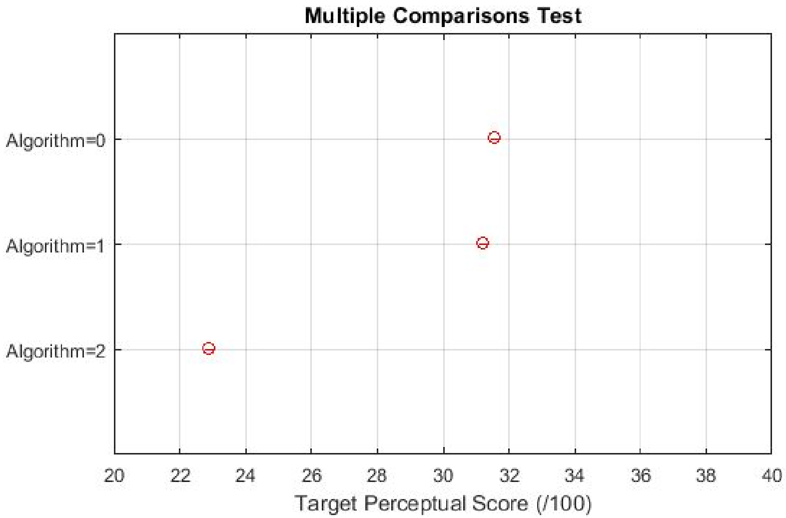
- Difficult to answer, have to make a qualitative judgement

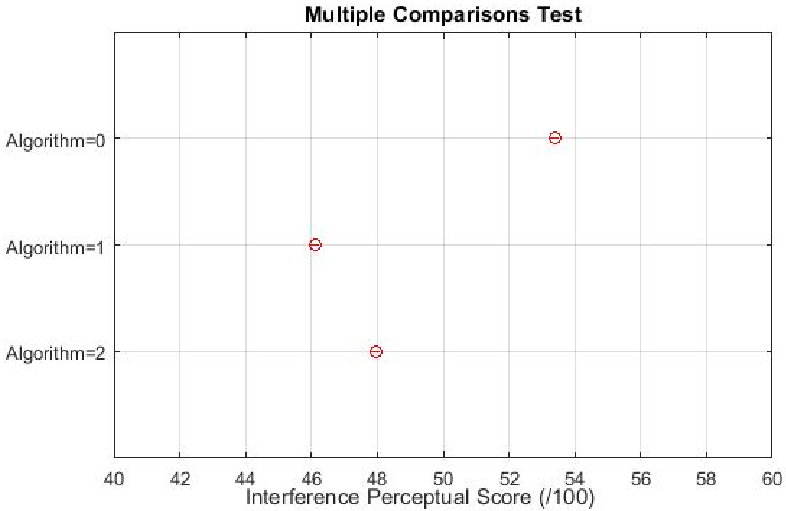
Lets Listen

Does the Implementation of the variants work?

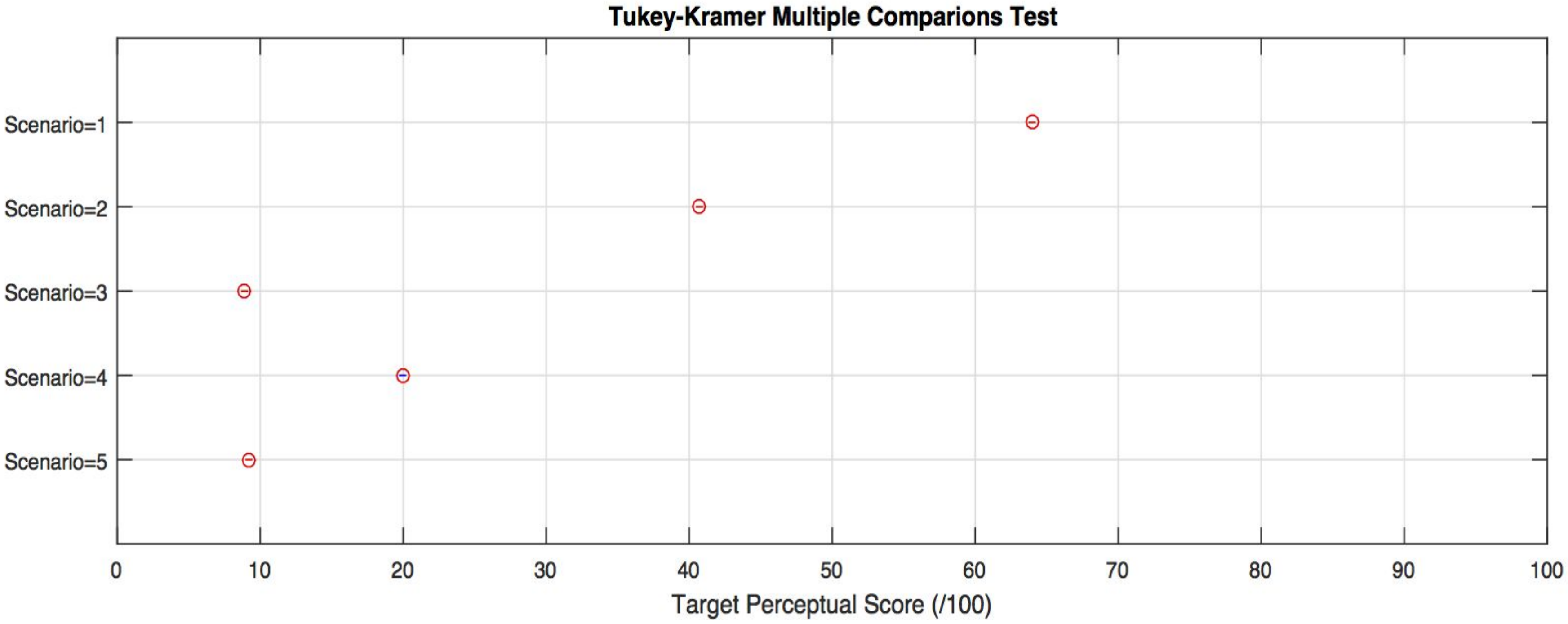
- Even more difficult to answer. Nothing to compare to, so must be compared relative to supervised variant.

- Semi-Supervised Target-Fixed Variant performed about as well as supervised variant

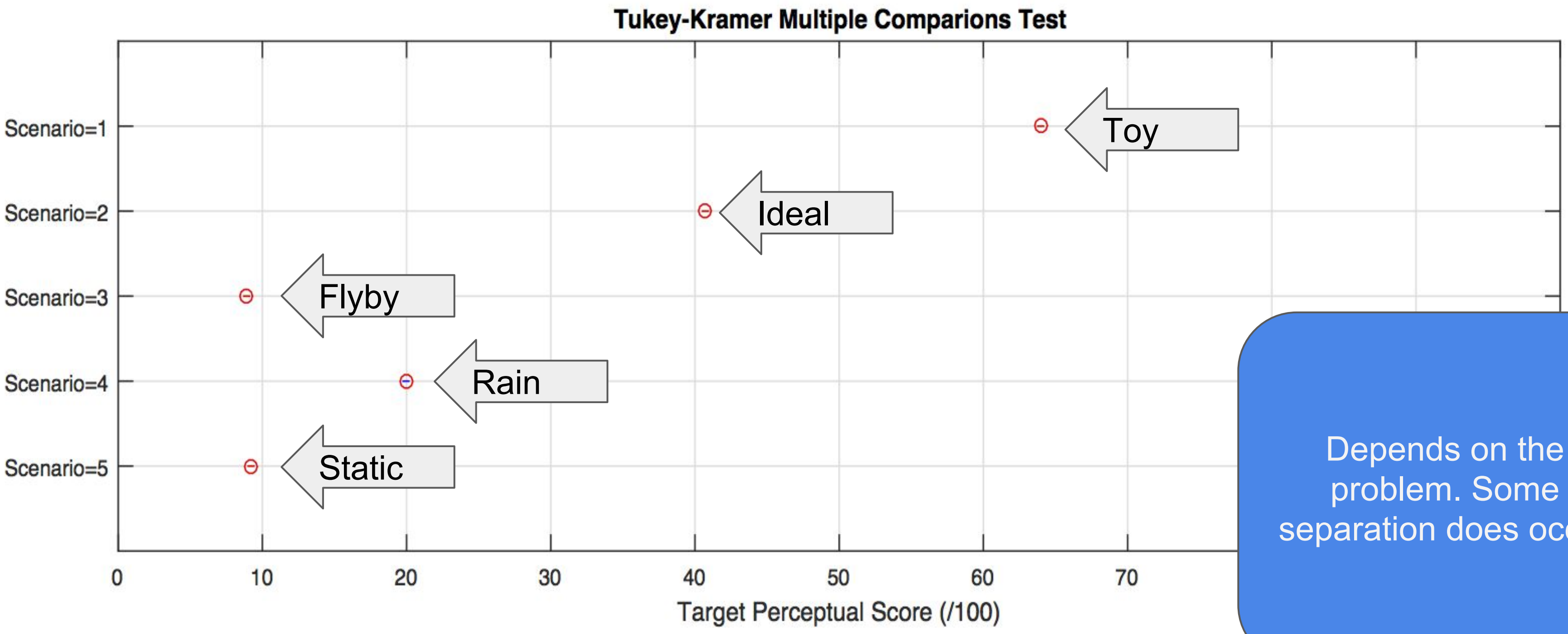




Is it Effective for Beehive Audio?



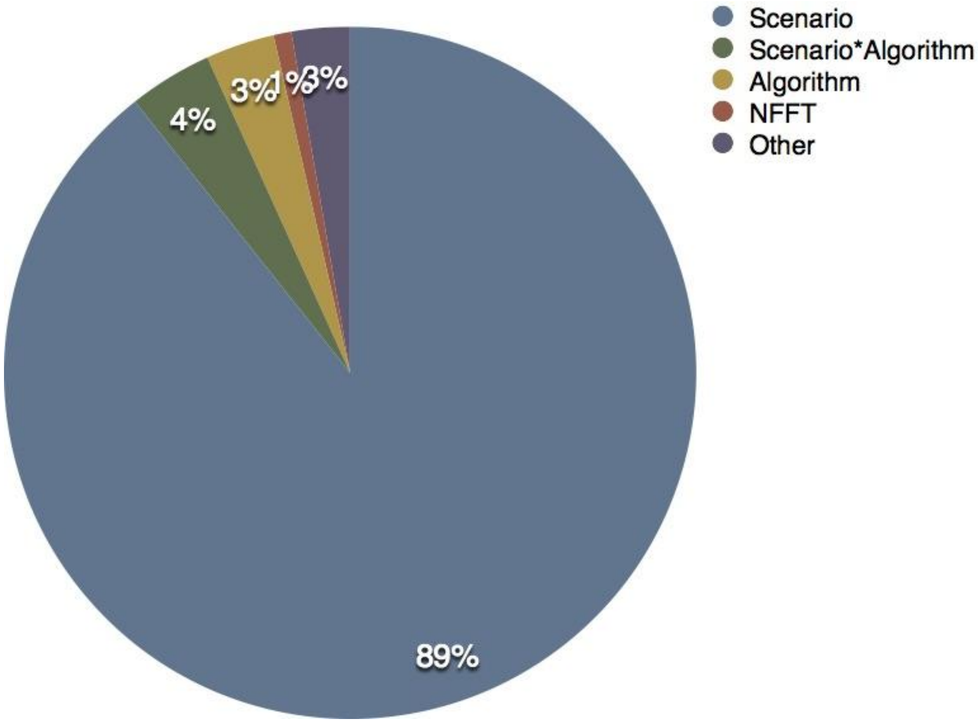
Is it Effective for Beehive Audio?



Depends on the problem. Some separation does occur

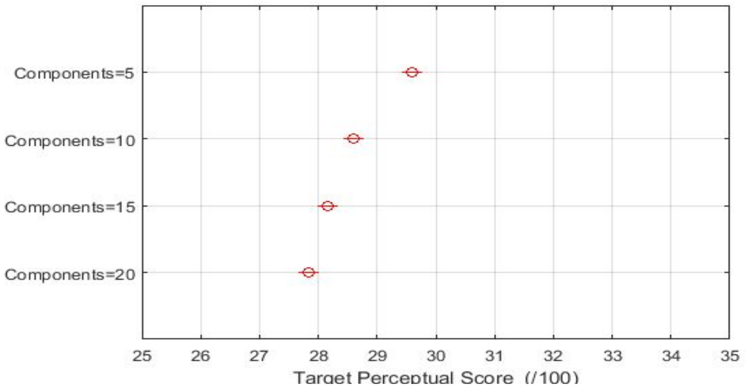
Optimize Parameters for Separation Quality?

- Difficulty of problem is the biggest factor
- We can still optimize our parameters



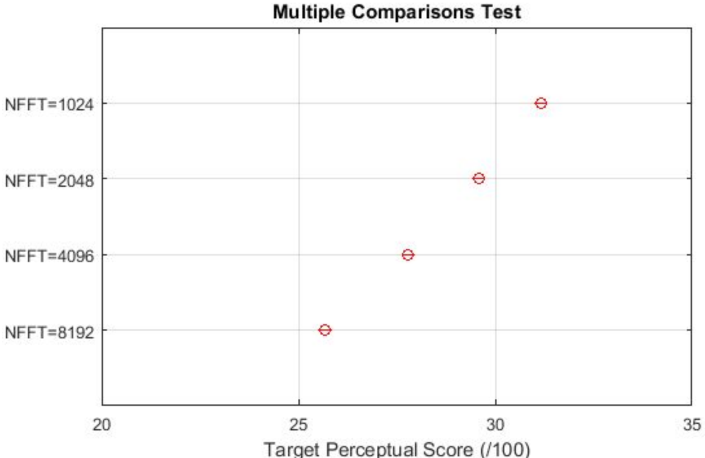
Optimize Parameters for Separation Quality?

#Components



#Components accounts for 0.0882% of total variance

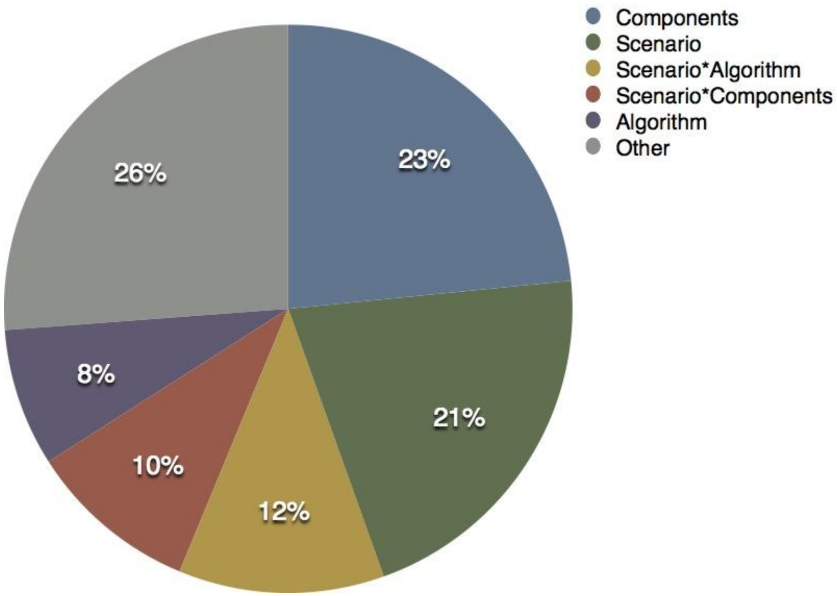
NFFT



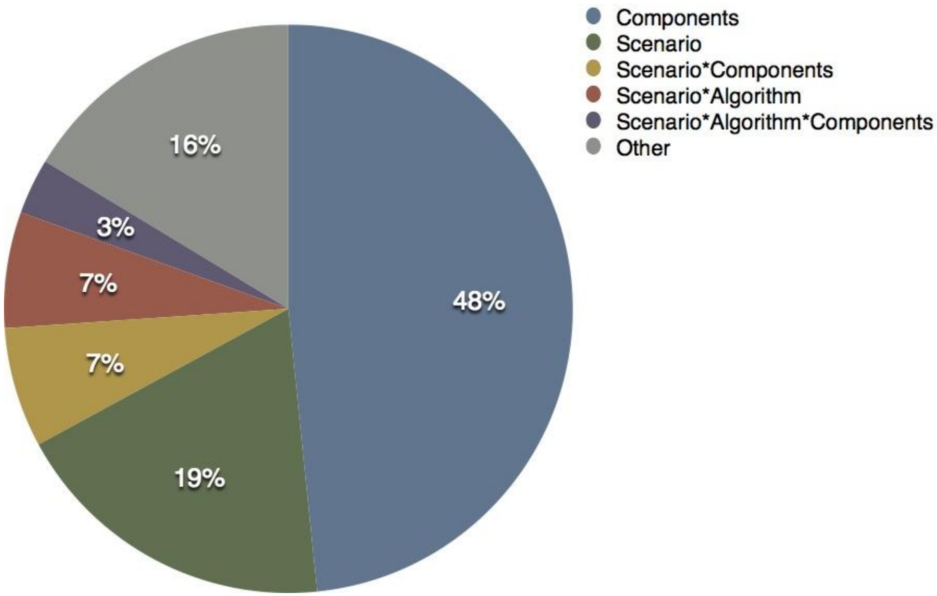
NFFT accounts for 0.829% of total variance

Optimize Parameters for Computation Duration

Training Phase Duration

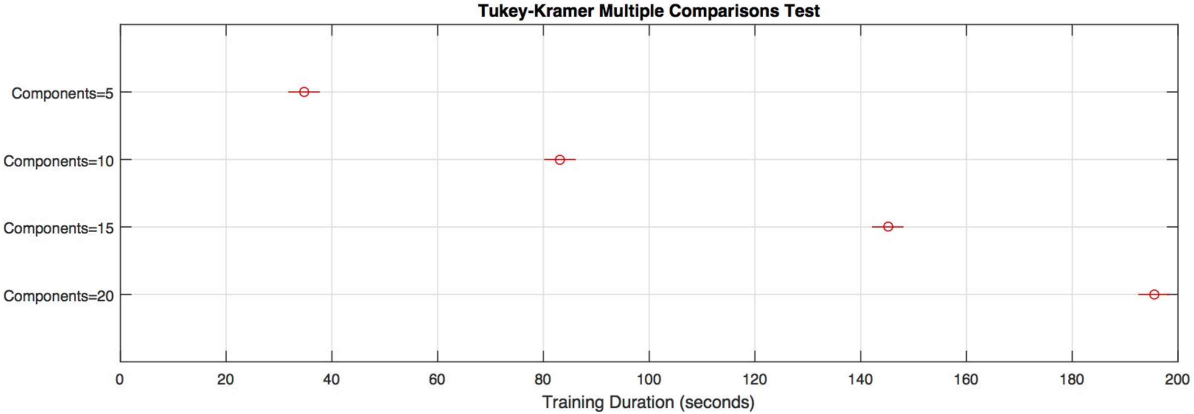


Separation Phase Duration



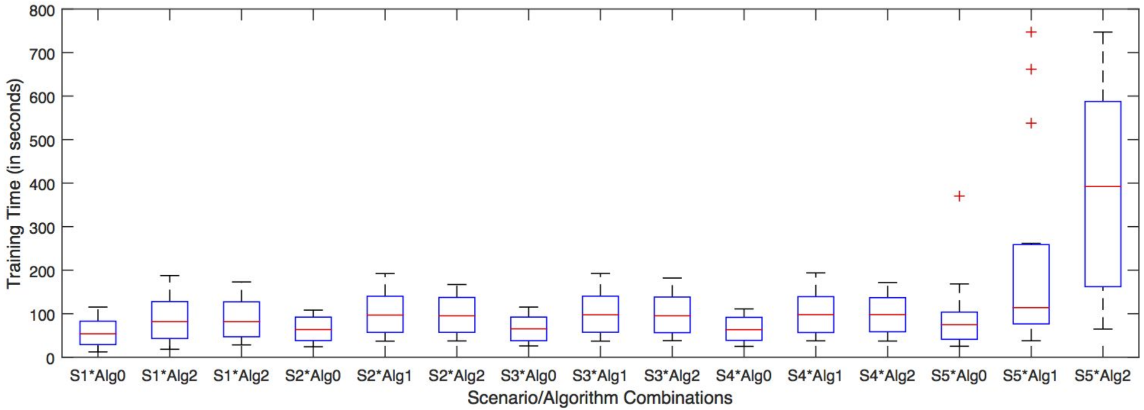
Optimize Duration of Training Phase

#Components



#Components accounts for **23.4%** of variance

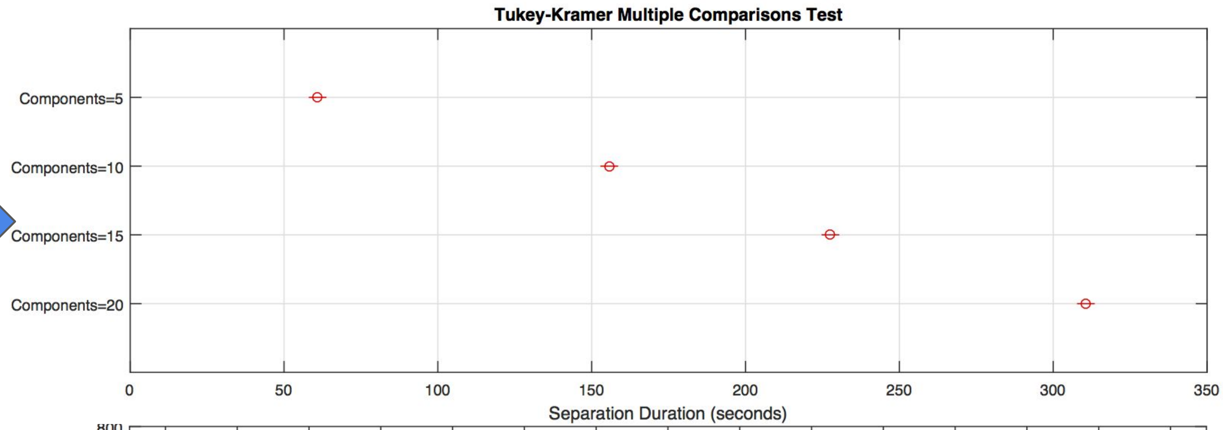
Scenario & Algorithm



Scenario and Algorithm combined accounts for **40.64%** of variance

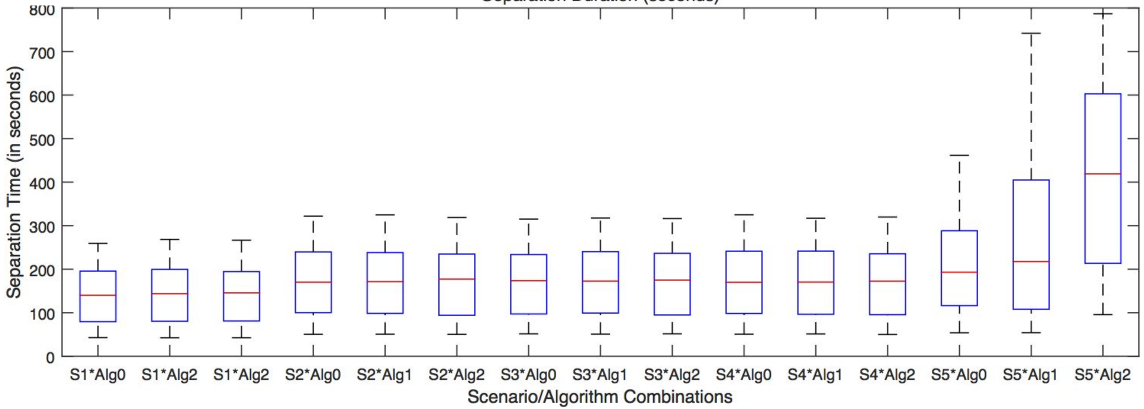
Optimize Duration of Separation Phase

#Components



#Components accounts for **48.4%** of variance

Scenario & Algorithm



Scenario and Algorithm combined accounts for **26.92%** of variance

Summary of Results

- Dependent on Problem
  - Similar two sources are more difficult they are to separate
  - Similar sources appear less likely to converge
- Use the supervised or semi-supervised fixed variants when possible
- As the number of components goes down, the performance improves
- As the NFFT goes down, performance improves
  - Time resolution more important than frequency resolution
- As the number of components goes up, computation duration increases

Limitations

- We still don’t understand the upper and lower bounds of each component
- Only 4 seconds of audio used for each sample
- We don’t have perfectly isolated sources
- How do we know that PEASS measures correlate to the fitness of a signal for analysis?

NFFT	**#Components**
...Lower?	...Lower?
1024	5
2048	10
4096	15
8192	20
...Higher?	...Higher?

?

BETTER

?

Future Work

- What are the upper and lower limits of each parameter?
- How does increasing training time impact performance?
- A new method of evaluation that is correlated with fitness for analysis
- Improve the Unfixed-Target variant
- More types of scenarios!

Conclusion

We outlined the **problem of CCD** and beehive audio

We explored **Latent-Variable Decomposition** as a possible solution

We posed some specific **research questions**

We **devised an experiment** to answer those questions,

We found that this technique shows **Promise** for certain problems. And now that the groundwork has been laid, future research can improve this technique

Acknowledgements

- Dr. Parry
- Dr. T
- Dr. Parks
- Dr. Smaragdis and Dr. Raj
- App State CS!