Modeling Blind Single Channel Sound Separation Using Predictive Neural Networks

Candidate #1002072
MSc in Neuroscience, Project 2 Dissertation, Trinity Term 2016

Word Count: 10,066 words

Abstract

The hum of an individual's morning commute may consist of a jackhammer from nearby construction work, the screeching wheels of a train, conversations of other commuters, and the audiobook they are currently listening to. Humans – and more broadly, all hearing animals – only receive the mixture of the sounds surrounding them, yet they unconsciously separate out and distinguish its individual sources. Although humans have two ears, they are quite capable of such source separation using one ear (i.e. a single channel). Ongoing neuroscience and engineering research on the problem of single-channel source separation has primarily focused on untangling sound mixtures into their individual components in a supervised fashion, in which both mixed and single sound sources are learnt from. However, in nature, sound separation is an unsupervised task, as creatures never have access to cleanly, separated-out sounds. This dissertation is the first known attempt to tackle this fuller problem of unsupervised single-channel source separation without hand-crafted algorithms. It presents a promising, novel unsupervised sound source separation paradigm by first training an artificial neural network to predict future cochleagram activity and then leveraging regularized network hidden units to isolate an individual component sound from a mixture. Additionally, this work identifies two biologically-consistent features in its networks: First, it shows that network hidden units exhibit frequency and temporal tuning comparable to those found in the spectro-temporal receptive fields (STRFs) of auditory cortex (AC) of anesthetized ferrets (Singer et al., n.d.). Second, it identifies hidden units whose selectivity mimics those of selective ferret AC neurons (Harper et al., n.d.). This dissertation suggests that predictive coding may not only encode biologically-consistent qualities but also capture features necessary for sound separation.

Table of Contents

Table of Figures

Table of Equations

Introduction

In many auditory situations, like a crowded conference room or a college gathering, there are multiple sounds from different sources occurring at once. Somehow, the brain is able to separate out these mixtures of sounds into separate sources, at least to the extent that sound recognition and identification can often occur. Examples include being able to isolate and understand a speaker from a background of other speakers, or hear a singer from amidst musical instruments, or distinguish a crackling fire in a hearth from the pouring rain outside. This challenging problem has been likened to identifying and characterizing the boats on a lake (i.e. different sound sources) from just observing the ripples in two small channels from the lake (i.e. our two ear drums) (Bregman, 1990). Although much progress has been made on computational solutions to the source segregation problem for some restricted situations, for the auditory situations typically experienced by humans and other animals, much remains to be solved.

This project had two aims with regard to this problem:

1. To investigate possible neurally-inspired computational models that can solve the problem subject to many of the constraints that the brain often faces.
2. To investigate whether the units in such models display similar properties to single neurons recorded in the primary auditory cortex, in response to isolated sounds and sound mixtures.

Improved understanding of the computational basis of sound source segregation would have many applications in engineering from automatic speech recognition in noisy environments, to enhancing auditory prosthetics, and may also have implications for source segregation beyond the auditory domain. Understanding the neural instantiation of sound source segregation would also be valuable in helping those with hearing difficulties, who often struggle in noisy environments (Kochkin, 2002), as well as possibly provide general insight into the manner by which the brain partitions the experienced world.

The blind source segregation problem as faced by the brain

The blind source segregation problem faced by the brain is extremely challenging. Most current computational methods for source segregation apply best in certain restricted situations, often unlike the auditory situations the brain has to cope with.

Here are just some of the major challenges in the source segregation problem encountered by the brain:

1. The brain has only one or two sound receivers (i.e. two ears). Humans can do sound segregation on monaural recordings, as most 20th century music were recorded using a single channel. Yet, independent component analysis (ICA), one of the main sound segregation paradigms, often requires as many microphones as there are sound sources (Comon, 1994).
2. The brain never receives the sources in a mixture, ruling out supervised models that dominate the sound segregation landscape.

3. The number of sources is rarely unknown and can vary with the passage of time, yet many models require there to be a fixed, known number of sources.
4. Natural sounds are extremely diverse: some are harmonic, while others are not; some are regular, while others are stochastic. However, some models rely on specific qualities, such as harmonic rules, that are only contained in a subset of natural sounds.

Previous Computational Approaches to the Source Segregation Problem

Until recently, computational approaches to sound separation have largely fallen under two frameworks: computational auditory scene analysis (CASA) and independent component analysis (ICA). Introduced in the early 1990s as the first formidable attempt on sound segregation (Bregman, 1990; Brown and Cooke, 1994), CASA leverages the harmonic structure of pitch-based noises to design hand-crafted algorithms for isolating individual sounds from their mixture (Kashino and Tanaka, 1993; Ellis, 1994; Wang and Brown, 2006). Around the same time, the mathematically-grounded ICA framework (Comon, 1994) and several implementations of it (Bell and Sejnowski, 1995; Cardoso and Laheld, 1996) was developed. An extension of principal component analysis, ICA treats sound separation as a matrix decomposition problem and seeks to find the weights that best explain how sounds additively combine to form a mixture. When there are as many microphones as there are sounds, the typically underdetermined problem can be solved and ICA can use differing distances to microphones to perform unmixing (Comon, 1994). More recently, ICA has been retooled for single-channel sound separation by masking and weighting different parts of the frequency spectrum (Roweis, 2001) and first learning the statistical qualities of different kinds of sounds (Jang and Lee, 2003).

However, when attempting to understand how nature solves sound separation, these main frameworks are severely lacking in that CASA requires strong assumptions that do not hold for atonal sounds like the crunching of leaves, while ICA does not consider the biological mechanisms of sound separation.

Previous Investigations on the Neural Basis of Sound Source Segregation

Recent experimental research using multi-electrode (Mesgarani and Chang, 2013) and EEG (O'Sullivan et al., 2015) recordings has shown that the auditory cortex neurons of humans cued to attend to a single sound in a mixture neurons are selective for and can be used to decode the target sound. Additionally, work on the auditory pipeline suggests that the more upstream a neuron of an attending human is in the auditory pathway, the more selective it is (Zion Golumbic et al., 2013). In animals, one paper demonstrated that some primary auditory cortex neurons of anesthetized cats were selective for the background noise of bird chirps, even when presented simultaneously with the louder, main chirp sound (Bar-Yosef and Nelken, 2007). Other related animal studies on auditory streaming (Kanwal et al., 2003; Micheyl et al., 2005; Bee et al., 2011) are quite limited in that they mostly used simple tones and focused on separating out sequentially-played "ABAB" tones rather than simultaneously occurring tones.

Our Approach to Sound Source Segregation and its Consistency with Neural Data

The approach taken here, which may have some potential to overcome some of the above challenges, is based on the idea of predictive coding. There are a number of different takes on what is meant by predictive coding, the study here takes the approach that neurons and the networks in which they are embedded are optimized to represent that which is predictive of the future (Bialek et al., 2001; Salisbury and Palmer, 2015). It is posited that if such network is constrained to be parsimonious, perhaps with some additional constraints, it may learn to naturally perform stream segregation as that is the most efficient way to accurately predict the future from the past.

This dissertation presents a novel paradigm for unsupervised source segregation by first training artificial neural networks to perform a supervised predicting the future task and second leveraging subsets of network hidden units to perform unsupervised sound separation. It also outlines a biologically-inspired network regularization technique that yields surprisingly decent unsupervised sound separation results.

In addition to presenting a promising avenue for unsupervised source segregation, this work also identified biologically-consistent auditory receptive fields and selective hidden units in its trained neural networks; the networks' hidden units demonstrated surprising similarity to characteristics of primary auditory neurons in anesthetized ferrets (Harper et al., n.d.; Singer et al., n.d.).

Methods

All computational methods were created using custom MATLAB and python code and utilized the NumPy (van der Walt et al., 2011), SciPy (Jones et al., n.d.), Matplotlib (Hunter, 2007), Lasagne (Dieleman et al., 2015), and Theano (Al-Rfou et al., 2016) Python libraries for numerical computing, statistical testing, visualizing results, and designing and training neural networks respectively.

Data

2.5-second, unique sound clips of individual, human speakers talking were taken from a database of over 532 Australian English speakers (301 female, 231 male) participating in one of three tasks: a casual telephone conversation, an information exchange task over the telephone, and a pseudo-police-style interview (Morrison et al., 2015). The sound clips were randomly assigned into two, equally-sized groups, A and B, and sound clips of individual speakers were pair-wise mixed together to form a third group, AB.

It was reasoned that a suitable database would contain mixtures of sounds that can mostly be segregated by the human ear; thus, only triplets of sound clips from A, B, and AB were used when they met the following criteria:

1. No more than 15% of either single-speaker clips from A or B was silent (i.e. zero amplitude)
2. The ratio between the root mean square (RMS) of the individual sounds must be greater than $4/3$
3. The RMS of either single-speaker sounds must be greater than 0.02

The RMS metric captured the overall amplitude of a sound and is given by Equation 1; thus, the second criterion ensured that there was enough of an average volume difference for the two speakers to be differentiated between and the minimum RMS criterion made sure that both sounds were sufficiently loud enough to be intelligible.

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {d_i}^2}$$

Equation 1: Root Mean Square (RMS) Metric

The above criteria thresholds, along with the 2.5-second clip length, were determined by a human listener on a sample of the sounds to enable the mixture sounds to be "unmixable" by the human ear in most cases, that is, that a human would be able to parse out the two individual sounds. In total, the above criteria filtered over 100,000 2.5-second single sound clips down to a set of over 16,000 single sounds to yield 8,268 triplets of mixtures AB and their component single-sounds A and B.

Each 2.5-second clip in the three groups A, B, and AB, was then represented as a cochleagram. Each clip was sampled at 41 kHz and converted into log-spaced cochleagram,

which corresponds to sound activity after its been processed in the cochlea. To generate the cochleagram, the power spectrum was taken using 50 millisecond Hanning windows, each overlapping by 25 ms, to yield 25 ms temporal resolution. This was done for 39 frequency bins evenly spaced by a sixth of an octave between $100 \times 2^0 = 100$ Hz and $100 \times 2^{6+1/3} \approx 8063$ Hz, which includes the typical frequency range for human speech (Figure 1). A whole 2.5-second cochleagram was then split up into 250-millisecond windows, each overlapping by 225 milliseconds, that is, shifted by 25 ms. In the prediction task, described in more detail below, a network was trained to predict the last 3 time bins, i.e. 75 ms, from the first 7 time bins, i.e. 175 ms, of a 250-ms window.

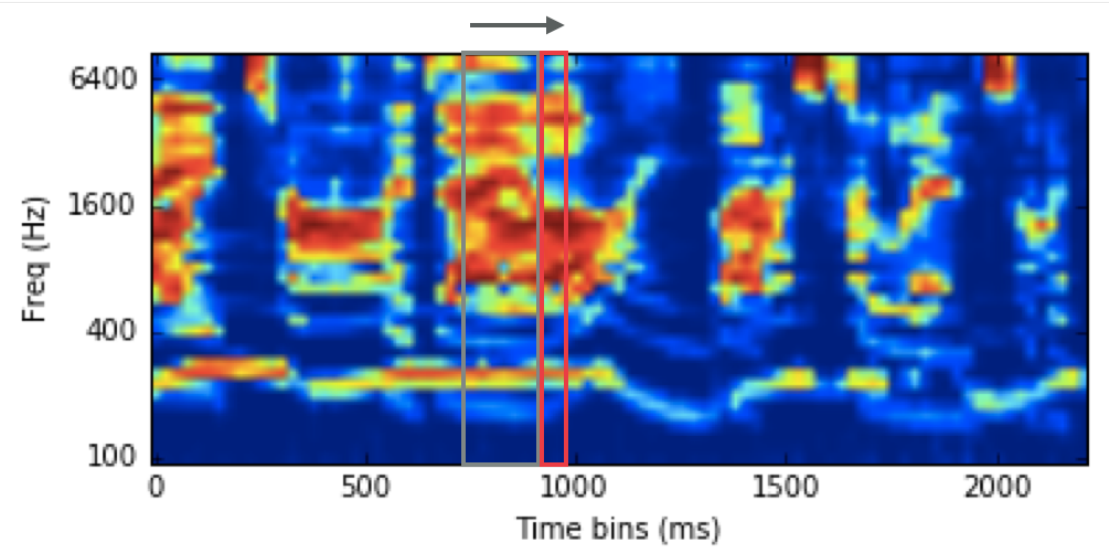


Figure 1: Cochleagram

2.5 second cochleagrams of human speech served as the inputs and outputs to the network models. Each cochleagram was sliced into 250-ms windows that overlapped by 225 ms. The first 175 ms of a cochleagram (grey box) was used to predict the next 75 ms of activity (red box).

Predicting the Future

The triplets of A, B, and AB sounds were split into training, validation, and test sets; 10% of cochleagram data was held out as a test set[1], and of the remaining 90% of data, 10% was held out as a validation set. In the training set, half of the cochleagram examples were from AB mixture sounds and the other half were from A and B single sounds that did not comprise the mixtures included in the training set, so that the network would not accidentally "memorize" and "associate" mixture sounds with their exact individual component sounds. In the validation set, all cochleagram examples from the triplets of validation A, B, and AB sounds were included. There were 8,640 2.5-second cochleagrams, which were then divided into 250-ms windows, in the training set and 2,160 2.5-second cochleagrams in the validation set. The networks were simply trained to predict the next 75 milliseconds from the previous 175 milliseconds of a same cochleagram window.

[1] The test set was not used in this dissertation because further work on this project will be done, so the test set is being saved for end-of-project testing.

For notation purposes, let X_{tr} be the set of input vectors of the training set, y_{tr} the set of target outputs of the training set, X_{val} the validation set's inputs, and y_{val} the validation set's target outputs.

Data Compression and Normalization

After the training and validation sets have been created, a series of data normalizations were applied.

First, all parts of all datasets – $X_{tr}, y_{tr}, X_{val}, y_{val}$ – are normalized in each of the 39 frequency channels by being divided by the median activity in that frequency channel in X_{tr}. Second, all parts of all datasets are compressed using the Hill function ($\alpha = 0.02$, Equation 2), which has been shown to model compression in the auditory nerve well (Hill, 1910; Lütkenhöner, 2008; Heil et al., 2011). The Hill function is mostly linear except at the extremes, where it has a biologically consistent compression effect on high intensities.

$$h(x;\alpha) = \frac{\alpha x}{\alpha x + 1}$$

Equation 2: Hill Function

Third, all parts of all datasets are z-score normalized by the mean, μ_{tr}, and standard deviation, σ_{tr}, of the training set inputs, X_{tr} (Equation 3).

$$z(x;\mu,\sigma) = \frac{x - \mu}{\sigma}$$

Equation 3: Standard (z-score) Normalization

Networks

Two kinds of artificial neural networks were trained using Lasagne (Dieleman et al., 2015)and Theano (Al-Rfou et al., 2016) python libraries: feed-forward, fully connected (FC) networks (FCNs) and recurrent neural networks (RNNs). For both kinds of networks, each layer is fully connected to the next layer (Figure 2); however, the hidden layers of RNN are also fully, recurrently connected to itself (Figure 4). The inputs to the networks were the first 175-ms slices of ordered, consecutive 250-ms cochleagram windows and the target outputs were the remaining 75-ms slices of the same windows (Figure 2).

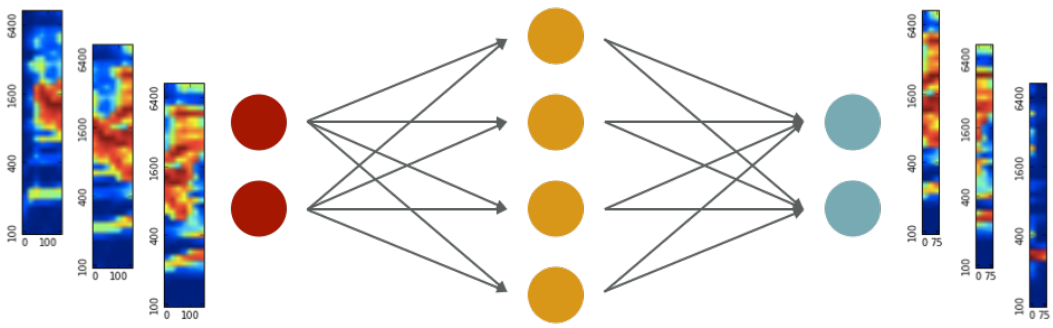


Figure 2: Diagram of 4-Hidden Unit Fully Connected (FC) Network with Cochleagram Inputs and Outputs

A benefit of an RNN is that its recurrent connections enable the network to "remember" previous inputs in a sequence by carrying over memory from the previous pass-through of the hidden layer. This feature is a loose analog to short-term memory; furthermore, it is known that the mammalian brain is not strictly feed-forward and incorporates recurrent connections (Douglas et al., 1995). In a predicting the future task, it is hypothesized that the more biologically-consistent recurrent connections of RNNs may yield sparser, more compact network representations and improved performance. One hidden layer with 100 hidden units was used for both kinds of networks.

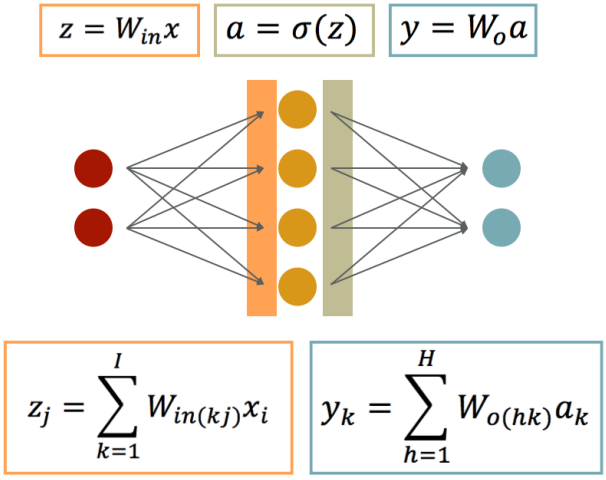


Figure 3: Model of 4-Hidden Unit Fully Connected Network (FC)

Top Row. *The equations for z, the vector of inputs to the hidden layer, a, the vector of outputs from the hidden layer after the non-linearity σ is applied, and y, the vector of outputs from the network that represents its predicted cochleagram of future sound activity, given in vector notation.* ***Bottom Row****. The equations for z_j, a scalar of the input to the j-th hidden unit, and y_k, a scalar of the k-th dimension of the network's predicted output, given in non-vector, summation notation.*

Mathematically, the FC (Figure 3) and RNN (Figure 4) models simply consist of a non-linear function σ, which is applied to their hidden units, as well as several bias terms (these are wrapped up in the weight matrices and not shown for simplicity in Figure 3 and Figure 4) and weight matrices:

1. W_{in}: an input weights matrix that transforms the input into its pre-non-linearity hidden unit representation,
2. W_o: an output weights matrix that transforms the post-non-linearity hidden unit representation to an output,
3. W_r: a recurrent weights matrix for RNNs only that incorporates the previous time step's hidden unit representation into its current one.

For RNNs, W_{in} and W_r were initialized by sampling uniformly from $[-0.01, 0.01]$. W_{in} for FC models and W_o for all models were initialized using the Glorot uniform initialization (Glorot and Bengio, 2010), which calls for sampling uniformly from $[-a, a]$, where

$$a = \sqrt{12/(fan_{in} + fan_{out})}$$

Equation 4: Glorot Uniform Initialization Term

fan_{in} denotes the number of incoming connections, and fan_{out} denotes the number of outgoing connections. All bias terms were initialized to zero-filled vectors of appropriate length.

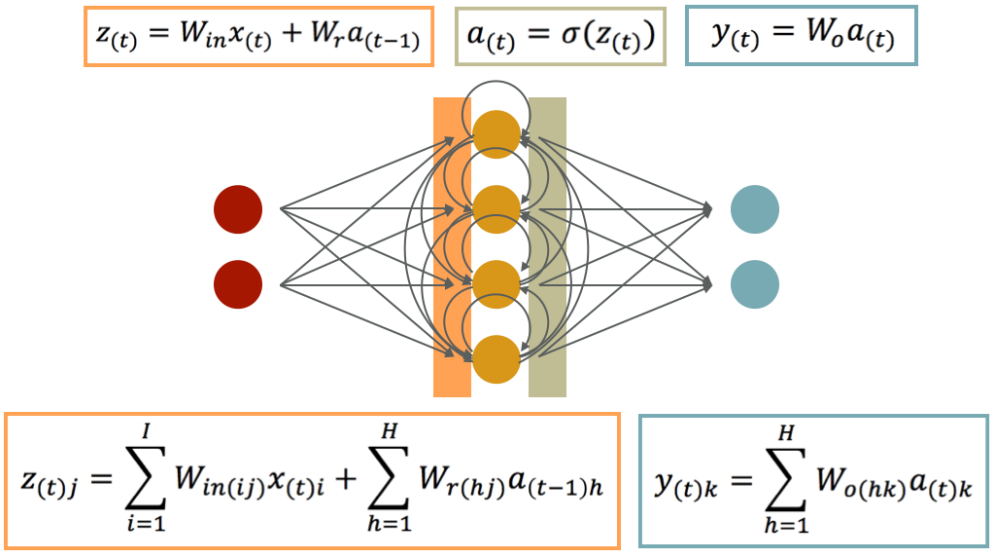


Figure 4: Model of 4-Hidden Unit Recurrent Neural Network (RNN)

The training of a neural network consists of two phases: 1., a forward pass through the network to compute outputs from the training inputs, and 2., a backwards pass to update the network parameters, θ, by minimizing a cost function.

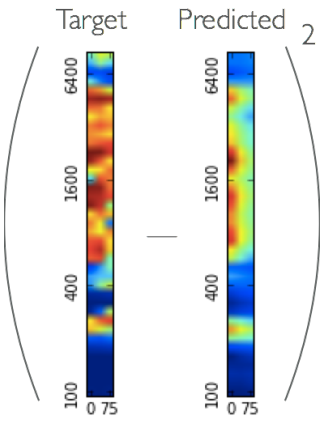


Figure 5: MSE Example

A mean squared error (MSE) cost function was used to train these networks (Equation 5). MSE effectively seeks to minimize the average difference between a computed output vector, $\hat{y}_{nt}$, and a target output vector, y_{nt}, where n denotes the training example and t denotes the instance in a sequence – in this case, the sequence the network is processing is the sequence of overlapping, cochleagram windows for a given 2.5-second sound (Figure 5).

$$\theta^* = \text{argmin}[\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \| y_{nt} - \hat{y}_{nt} \|_2^2 + \lambda \| \theta \|_1]$$

Equation 5: Mean Squared Error (MSE) Loss with L1 Regularization

To promote sparse compact network representations, L1 regularization was also applied (Equation 6a), which penalizes unnecessarily large network parameter values (Equation 5).

$$\| x \|_1 = \sum_{d=1}^{D} |x_d|$$

$$\| x \|_2 = \sqrt{\sum_{d=1}^{D} (|x_d|)^2}$$

Equation 6a,b: L1 and L2 Regularization of Vector x

Hyper-parameter Search

The performance of a neural network for a specific task and dataset is highly dependent on finding the right combination of network hyper-parameters. The basic neural networks trained here had three hyper-parameters: 1., the update function used to update network weights based on the cost function, 2., the non-linear function to apply to the hidden units, and 3., the L1-regularization parameter, λ, to scale the regularization penalty.

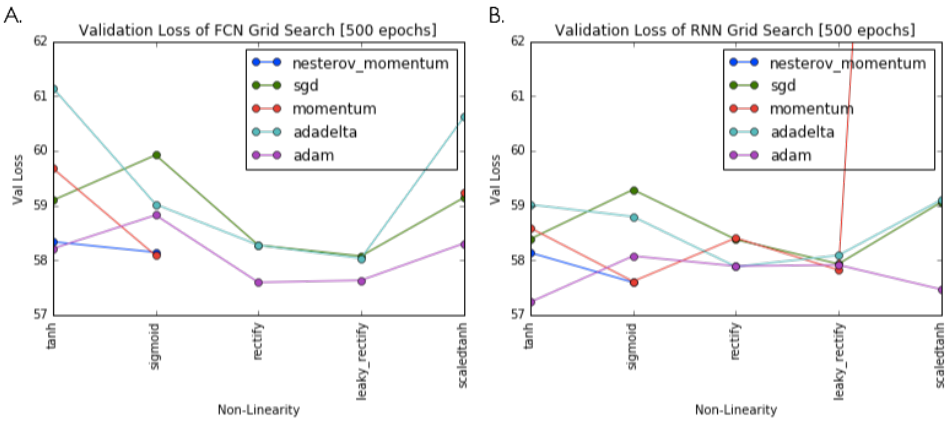


Figure 6: Hyper-parameter Grid Searches for Update and Non-Linearity Functions

*The typical FC model with the best predictive loss on the validation set used an adam update function and a rectify non-linear function (**A**), while the typical RNN model with the best loss validation used an adam*

*update function and a tanh non-linearity (**B**). RMSProp and Adamax were excluded from these figures because they yielded poor losses and/or unstable results.*

After some preliminary exploration, the L1-regularization parameter was set to $\lambda = 10^{-4}$. A hyper-parameter grid search over several options for update functions and non-linearities was performed to find the optimal update function and non-linearity pairing that best minimized the error on the validation set. The grid search identified that a FC network with an adam update function (Kingma and Ba, 2014) and a rectifier non-linearity (Equation 7) yielded the best predictive loss on the validation set out of all other combinations for FC models, while an adam-tanh RNN yielded the best loss of RNN models (Figure 6).

$$\sigma(x) = \max(0, x)$$

Equation 7: Rectify function

All networks described in this dissertation were trained for 500 epochs, meaning that the training set was iterated through 500 times, in which each iteration was followed by an update of network parameters (i.e. bias terms and weight matrices). It was found that network performance converged around 500 epochs and that training for longer did not yield substantial improvements and at times led to network instability.

"Explicit Hidden Unit Populations" Network Extension for Unsupervised Source Separation

An "explicit hidden unit populations" extension to the typical definitions of FC and RNN models was made, in order to encourage different sets of hidden units to be selective to individual component sounds.

This novel extension defines three mutually exclusive sets that comprise the entire hidden unit population: "group 1" (G1), "group 2" (G2), and "helper" (H) units. Recall that half the training inputs are mixtures and the other half are single sounds. The hidden unit activity of the different population groups is then regularized so that, ideally, G1 units are always used to process a single sound, G2 units are only used or recruited to represent a second sound in a mixture, and H units are used for general auditory processing; this penalty was added to the loss function (Equation 8). This ideal situation was encouraged by 1., penalizing G2 activity more than G1 units and 2., enforcing L1 norm competition between the G1 and G2 populations, while having L2 norm regularization on the activity within each population. L2 regularization tends to cause hidden units to all be small values, while L1 regularization tends to drive to zero any unnecessary hidden units. Thus, it was hypothesized that regularization technique would cause units in a group to act together but would cause groups to compete with one another. Lastly, L1 regularization was applied to H units to encourage efficiency and their output connections were eliminated; thus, they are only useful in RNNs as recurrent connections in the hidden layer.

$$P = \alpha \parallel a_{(G1)} \parallel_2 + \beta \parallel a_{(G2)} \parallel_2 + \gamma \parallel a_{(H)} \parallel_1$$

Equation 8: "Explicit Hidden Unit Populations" Penalty

The networks that leveraged this extension were trained with the following population split: 40% G1, 40% G2, and 20% H units, with $\beta \geq \alpha$ and $\gamma = \frac{\alpha\sqrt{\frac{|G1|}{|H|}|U|}}{2|U|}$, where $|G1| = 40$, $|H| = 20$, and $|U| = 100$ denoted the number of G2, H, and all hidden units. γ was set as it was so that each of the H units would experience half the amount of regularization as the G1 units. Explicitly defining these hidden unit populations allowed for further analysis of how a predictive network could tackle unsupervised source separation.

Hyper-parameter Search for Hidden Unit Population Regularization Parameters, α and β

A hyper-parameter search was conducted to find the optimal α first and then β. Using the L1-regularization term $\lambda = 10^{-4}$ as well as the best update and non-linear functions found in the grid search using typically-defined networks (i.e. adam-rectify for FC and adam-tanh for RNN), the α-varying grid search for FC and RNN models with explicitly defined hidden unit populations searched the space of $\alpha \in \{10^{-10}, 10^{-9} \ldots, 10^{-2}, 10^{-1}\}$ while $\beta = 1.2\alpha$. With predictive loss being roughly equivalent across tested $\alpha \in [10^{-10}, 10^{-6}]$ for both FC and RNN models (Figure 15a and Figure 17a), $\alpha = 10^{-7}$ was chosen and used in the β^*-varying grid search across the space of $\beta^* \in \{2^0, 2^1, \ldots, 2^8, 2^9\}$ and $\beta = \beta^*\alpha$. From the β^*-varying search, the $\beta^* = 16$ FC model and the $\beta^* = 2$ RNN model earned the best predictive losses in their respective model classes (Figure 15b and Figure 17b).

Thus, the $\beta^* = 16$ FC model primarily and $\beta^* = 2$ RNN model secondly will be analyzed in-depth in the Results section. These model choices for in-depth analysis were chosen a priori based on their predictive loss and before any knowledge of their unsupervised source separation abilities.

Unsupervised Source Separation

Probing Network on Unsupervised Source Separation Problem using Select Hidden Units

Given a set of hidden units, the network can be probed using solely that set of hidden units, i.e. G1 or G2, by turning off the output connections from all other units (Figure 7). Ideally, when a mixture sound is passed into a probe network, with only a subset of output connections enabled, as an input, it outputs a prediction of one of its component sounds instead of a prediction of the mixture.

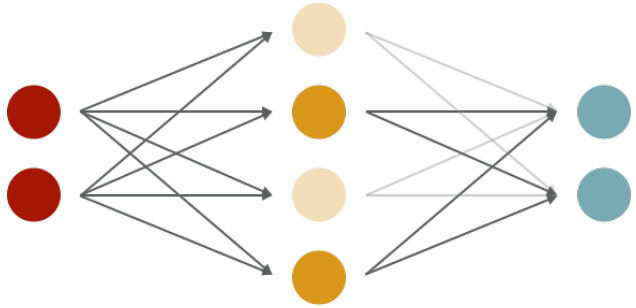


Figure 7: Example Network Using Only Output Connections from Hidden Units 2 and 4

Qualitatively, a probe network's predictions of individual triplet examples from the validation set can be visualized and examined to see whether the mixture's prediction is similar to the target prediction of one of its component sounds (Figure 12). Quantitatively, a measure of how well a probe network performs on the unsupervised source separation problem of predicting the output of a single sound from a mixture using only the output connections from a population of hidden units, i.e. $P = G1$, can be calculated by Signal-to-Distortion Ratio (SDR) measure (Equation 9):

$$E_{SDR}(P, X_m) = SDR(P, X_m) = 10 \log_{10}\left(\frac{\| y_{X_m} \|_2^2}{\| y_{X_m} - \hat{y}_{P(AB_m)} \|_2^2}\right)$$

Equation 9: SDR Metric for Probe Network's Source Separation

where y_{X_m} denotes the target output for the m-th individual sound $X \in \{A, B\}$ in the set of validation triplets and $\hat{y}_{P(AB_m)}$ denotes the output predicted for AB_m, the mixture sound of A_m and B_m by the network when using the output connections of the population of hidden units P. Intuitively, the SDR metric is positive when the signal of the target output, which is described in the log term's numerator, is greater than the distortion incurred by the prediction, which is captured in the log term's numerator. Thus, a larger SDR metric suggests a better source separation and a positive SDR score suggests that the prediction captures more signal than error.

For models with explicit hidden unit populations G1 and G2, because either G1 or G2 could be predicting either sound A or sound B for a given validation triplet, the SDR metric for the parallel pairing of G1 predicting sound A and G2 predicting sound B (G1-A, G2-B) must be compared to that for the cross pairing of G1 predicting sound B and G2 predicting sound A (G1-B, G2-A). To do so, the SDR metric can be computed for all combinations of hidden unit populations and single-speaker sounds, i.e., $E_{SDR}(P_1, A_m)$, $E_{SDR}(P_2, B_m)$, $E_{SDR}(P_1, B_m)$, $E_{SDR}(P_2, A_m)$ where $P_1 = \text{G1}$, $P_2 = \text{G2}$, and A_m and B_m denote sound A and B in the m-th validation triplet. Then, the scores for each pairing can be summed up (Equation 10b,c) and the maximum of the two summed scores would suggest the better sound-separating pairing. An overall SDR metric can be calculated by taking the mean of the maximum summed scores over all validation triplets, as given by Equation 10a:

$$L = \frac{1}{M} \sum_{m=1}^{M} \max(M_1, M_2)$$

$$M_1 = E_{SDR}(P_1, A_m) + E_{SDR}(P_2, B_m)$$

$$M_2 = E_{SDR}(P_2, A_m) + E_{SDR}(P_1, B_m)$$

Equation 10a,b,c: Overall Unsupervised Sound Separation Probe Measure using SDR Metric

Although the focus on this dissertation's

Results section will be on the analysis of network models with explicitly defined hidden unit populations, this probe network paradigm can be used on analyze any typical network if a set of hidden units with source separating qualities can be identified as well as generalized to separate out more than 2 sounds.

Unless otherwise noted, all visualizations of predicted cochleagrams (i.e. Figure 12) show the first 25ms – rather than the full 75ms prediction – predicted by a given network for all time bins over the 2.5 second period.

Statistical Testing

Permutation Test

To test whether hidden unit populations P_1 and P_2 (i.e. $P_1 = G_1$ and $P_2 = G_2$) used to probe the network produce source separations that are significantly better than if random hidden unit populations were used, a permutation test can be utilized. To generate a sample for the permutation test's null distribution, all hidden units with output connections are randomly split into two groups R_1 and R_2 that are then used to calculate probe predictions $\hat{y}_{R_1(AB_m)}$ and $\hat{y}_{R_2(AB_m)}$ for all $m \in \{1, \dots, M\}$ validation triplets. Next, an overall mean probe score can be calculated using $E_{SDR}(R_1, A_m)$, $E_{SDR}(R_2, B_m)$, $E_{SDR}(R_1, B_m)$, $E_{SDR}(R_2, A_m)$ and the max function (Equation 9). This mean score is then used as a sample. With a sufficiently large number of samples N, a p-value can be computed by calculating what proportion of samples in the null distribution are better probe scores than that earned by the tested populations P_1 and P_2. For instance, if $N = 10{,}000$, the p-value testing the null hypothesis that populations P_1 and P_2 do not produce better source separations than randomly assigned populations R_1 and R_2 is given by Equation 11:

$$p = \frac{|x_n > x|}{N}$$

Equation 11: p-value Calculation for Permutation Test using SDR Measure

where x is the overall SDR metric earned by populations P_1 and P_2, x_n is the n-th sample SDR metric, and $|x_n > x|$ denotes the number of samples that had a better source separation metric than one in question.

Inference Test using Correlation Coefficient and Student's t Distribution

To test the strength of a relationship between two variables, the sample Pearson correlation coefficient (CC), r, is computed. A two-sided p-value is also computed using the Student's t distribution to test the null hypothesis that, given the sample correlation coefficient r, the true correlation coefficient $\rho = 0$. The following t-value is calculated using Equation 12:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Equation 12: t-value for Correlation Coefficient Inference Test

where n is the number of pairs being compared. Then, the two-sided p-value can be calculated using the cumulative distribution function (CDF)[2], $F(x, df)$, of the Student's t distribution for $df = n - 2$ degrees of freedom (Equation 13):

$$p = 2(1 - F(|t|, n - 2))$$

Equation 13: Two-sided p-value for Correlation Coefficient Inference Test using the t Distribution's CDF

Scipy's linregress function (Jones et al., n.d.) was used to compute r, t, and p as well as to fit a linear regression line to the set of (x,y) coordinates of the two variables being compared.

Biologically-Consistent Characteristics

Fan-In and Fan-Out Units

To analyze how trained networks make their predictions and compare the "receptive fields" of the hidden units with spectro-temporal receptive fields (STRFs) of ferret auditory cortex neurons, the input and output network weights associated with a given hidden unit were appropriately reshaped and normalized per unit to show a given unit's temporal and frequency tuning (Figure 8).

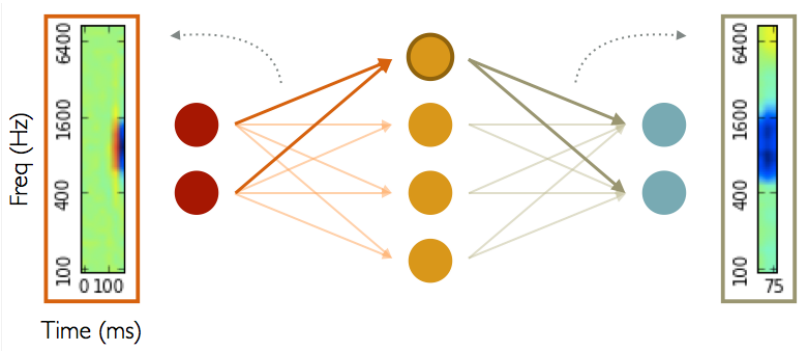


Figure 8: Example of How Fan-In and Fan-Out Weights for Hidden Unit 1 are Reshaped to Form a STRF Visualization

Selectivity Measures

To examine how "selective" a given hidden unit was to a single sound in a mixture, for each triplet set of sounds from sets A, B, and AB in the validation set, the Pearson sample correlation coefficients were calculated between the outputs of a given hidden unit (Equation 14a,b) for the A and AB sounds, as well as between the given hidden unit's outputs for the B and AB sounds.

$$a_j = \sigma(z_j) = \sigma\left(\sum_{i=0}^{I} W_{in(ij)}\, x_i\right)$$

$$a_{(t)j} = \sigma(z_{(t)j}) = \sigma\left(\sum_{i=0}^{I} W_{in(ij)}\, x_{(t)i}\right)$$

[2] For any distribution, the CDF function denotes the probability that a random variable X will be less than or equal to some sample x: $F(x) = P(X \leq x)$.

Equation 14a,b: Output of the j-th Hidden Unit for FC (top) and RNN (bottom) Models

Then, for each triplet, $(CC(A,AB), CC(B,AB))$ was plotted onto a Cartesian plot and a 2D histogram was used to visualize all the points (Figure 9).

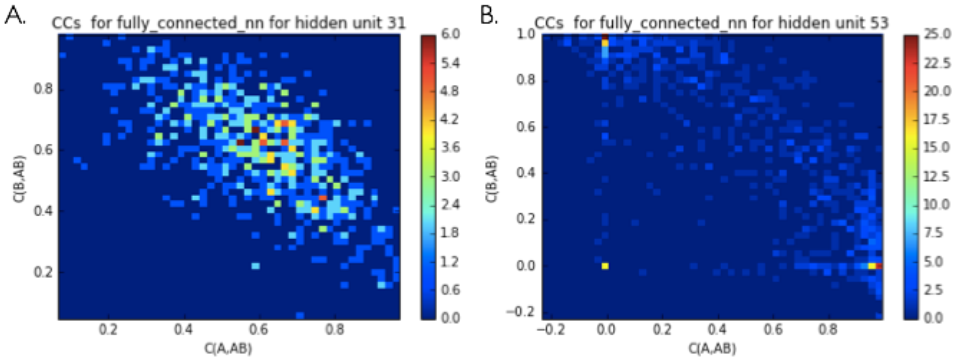


Figure 9: Example of Low (Left) and High (Right) Selective Hidden Unit Visualization

Qualitatively, points that lie along the x- or y-axis close to (1,0) or (0,1) suggest that the given hidden unit's response was selective for one sound in the mixture but not the other, because its mixture response was lowly correlated with that of one single sound but highly correlated with that of the other sound. In contrast, a point close to the x = y line suggests that the unit's response was non-selective for that triplet example because its mixture response is as equally correlated with one sound as it is with the other sound. To better quantify a unit's overall selectivity, the selectivity measure given by Equation 15 was calculated for each triplet example's point; a visualization of the selectivity index is given by Figure 10.

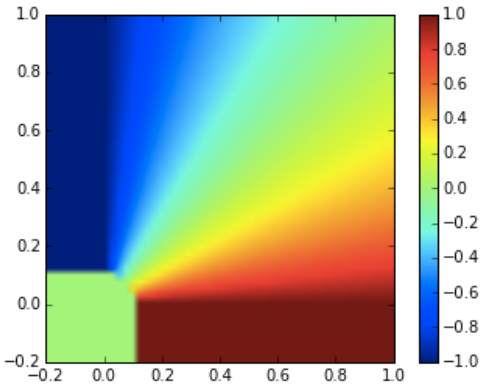


Figure 10: Selectivity Index

$$s = \begin{cases} 0, & \text{if } \max(0,x)^2 + \max(0,y)^2 < 0.01 \\ \dfrac{\max(0,x) - \max(0,y)}{\max(0,x) + \max(0,y)}, & \text{otherwise} \end{cases}$$

Equation 15: Selectivity Measure

Then, a highly selective unit (Figure 9b) whose visualization showed points along the x- and y- axes yielded a selectivity distribution with two large peaks at -1 and $+1$, while a less selective unit (Figure 9a) whose visualization showed points mostly near the $x = y$ line yielded a selectivity distribution with one large peak at 0.

A single, scalar aggregate selectivity measure was also calculated by computing the mean squared selectivity score for a given unit (Equation 16):

$$S = \frac{1}{M} \sum_{m=1}^{M} {s_i}^2$$

Equation 16: Aggregate Selectivity Measure for a Hidden Unit

where M denotes the number of triplet examples in the validation set and s_i denotes the selectivity measure for the i-th triplet example for a given hidden unit. The aggregate selectivity scores for all hidden units were then used to identify "least" and "most" selective units.

Experimental Methods

The reconstructed receptive field visualizations were compared to spectro-temporal receptive fields (STRFs) of primary, auditory cortex neurons recorded in anesthetized ferrets (Figure 19a). These recordings were conducted and first described in another work (Willmore et al., 2016). Willmore et al. used the BigNat dataset, which is comprised of all natural sounds, and described in detail how STRFs were derived from the recorded spiking data. In short, a mean squared error (MSE) linear regression between a neuron's spiking pattern over time and the log spectrogram of the corresponding auditory stimuli was conducted. Unlike Willmore et al., L1 regularization for the regression was used instead of L2. The resulting STRFs ranged over 32 frequency bins from 500 to 17,827Hz with one-sixth octave spacing and 40, 5ms time bins to be directly comparable to the hidden units of the neural networks in Singer et al., which were trained on 215ms auditory stimuli. In comparison, the auditory stimuli used to train the networks in this dissertation used 39 frequency bins ranging from 100 to 8093Hz, which is typical for human speech, and coarser, 25ms time bins, with which 7 time bins of 175ms stimuli were used to predict the next 75ms' activity.

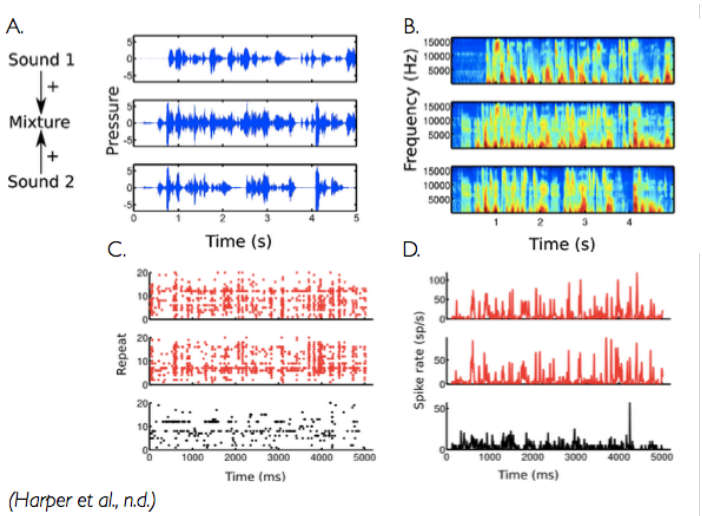


Figure 11: Data for Selectivity in Source Separation Experiments in Ferrets (Harper et al., n.d.)

Top Row. *The audio waves (**A**) and cochleagrams (**B**) of two natural sounds and the sound mixture they form.* ***Bottom Row. C.*** *Rasta plots, where each row represents a repeat of the stimulus and each dot represents a spike, of the responses a primary auditory cortex neuron of an anesthetized ferret has to two natural sounds (top and bottom row) and their mixture sound (middle row).* ***D.*** *The average peri-stimulus time histograms, which averages the spike rate over all repeats and bins the averages into 10ms time windows, corresponding to the plots in **C**.*

The selectivity visualizations of the hidden units of this dissertation's networks are also compared to similar ones of primary auditory cortex neurons of anesthetized ferrets recorded *in vivo* with multi-electrode probes (Figure 20). 16, 5-second audio clips of single natural sounds were used as stimuli, as well as 16 clips of mixture sounds that were each composed of two randomly selected sounds from the set of 16 single sounds (Figure 11a,b). The 32 sounds were played in random order and repeated 20 times for each penetration of the neural probe, in order to extract 20 per-neuron responses for each sound (Figure 11c,d).

Results

In this section, three findings are presented. One main finding:

1. a feed-forward, fully connected neural network that predicts future sound activity from past activity and has distinctly regularized hidden unit populations, which show capacity for unsupervised, single channel sound source separation,

and two minor findings:

2. the similarity between the network's receptive fields and real spectro-temporal receptive fields of ferret, auditory cortex neurons, and
3. the similarity between the network's hidden units' sound selective properties and that of ferret, auditory cortex neurons.

In addition to demonstrating the above novel, promising, and somewhat biologically-consistent paradigm for unsupervised source separation, the following other lines of research were explored but were not included in this dissertation:

1. a supervised approach to sound source separation using fully connected (FC) and recurrent (RNN) neural networks with a modified loss function that encourages predicted single sounds to be as distinct from each other as possible (Huang et al., 2015),
2. an unsupervised paradigm that involves identifying selective hidden units in typically FC and RNN models and leveraging them to perform source separation,
3. similar use of RNNs whose recurrent weight matrices were specially initialized as scaled versions of the identity matrix (Le et al., 2015), and
4. the use of several sound separation metrics based on comparing the target and predicted sound separation output using mean square error and correlation coefficients.

Unsupervised Source Separation

Without any explicit supervised training on source separation, a fully connected (FC) network model trained on the predicting the future of its input was able to perform with promising capacity on the unsupervised source separation problem. The model was trained using the following hyper-parameters:

- Update function: adam
- Non-linearity: rectifier
- L1 regularization factor on network weights: 10^{-4}
- Regularization factors for explicit hidden unit populations (G1, G2, H): $\alpha = 10^{-7}, \beta = 16\alpha, \gamma = \frac{\alpha\sqrt{\frac{|G2|}{|H|}U}}{2U}$

The selection of these hyper-parameters are described in the Methods section and were chosen before observing the network's source separation ability because they yielded near-best performance on the predictive task.

Figure 12 shows qualitative examples of the network predicting the separated out individual components of a mixture by only using the output connections of an explicit population of hidden units (i.e. G1 or G2). These two populations were regularized in training to both compete with one another as well as regularized differently so that G2 received 16 times the regularization penalty for its activity as G1, because $\beta = 16\alpha$. The examples were selected for earning the best Source to Distortion ratio (SDR) scores, where a higher score connotes a better separation result and a positive one denotes that the source separation contains more signal than error. It was observed that the hidden unit populations G1 and G2 could each capture – to an extent – the characteristics of primarily one of the individual sounds.

To test whether the networks with distinctly regularized hidden unit populations G1 and G2 were truly outputting decent, unsupervised source separation results, a permutation test was conducted (Supplementary Figure 23). Each of the 10,000 samples was generated by randomly splitting the network's hidden units with output connections into two populations and computing the SDR sound separation metric using those random populations. The model's SDR score when using the actual G1 and G2 populations was better than all but 350 of the 10,000 scores ($p = 0.035$) earned by randomly split hidden unit populations.

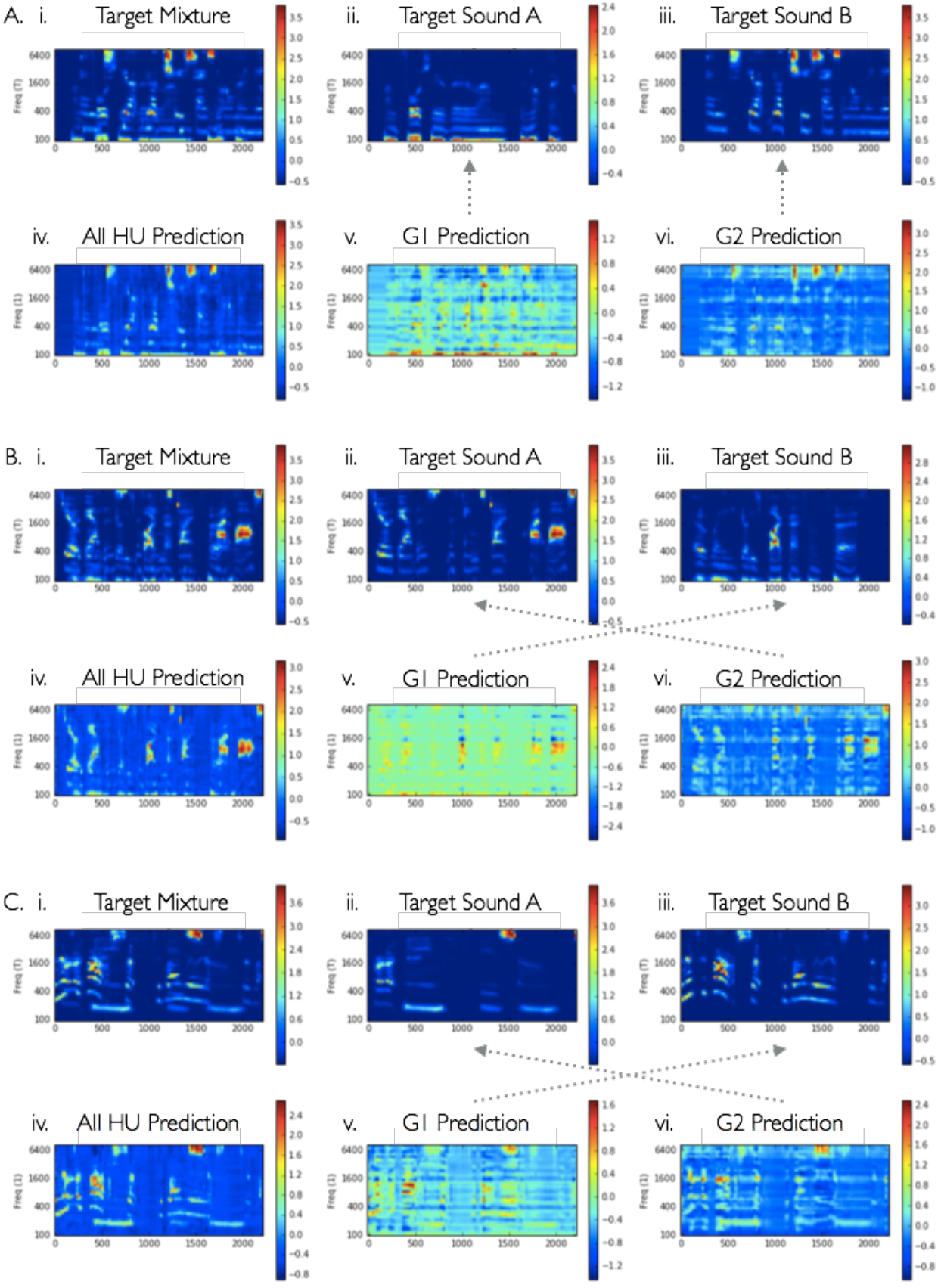

A. i. Target Mixture
ii. Target Sound A
iii. Target Sound B
iv. All HU Prediction
v. G1 Prediction
vi. G2 Prediction
B. i. Target Mixture
ii. Target Sound A
iii. Target Sound B
iv. All HU Prediction
v. G1 Prediction
vi. G2 Prediction
C. i. Target Mixture
ii. Target Sound A
iii. Target Sound B
iv. All HU Prediction
v. G1 Prediction
vi. G2 Prediction
Freq (T)
Freq (1)

Figure 12: Examples of Unsupervised Source Separation with the Best SDR Probe Scores

*For each example (**A, B, C**), the target cochleagrams for the mixture (**i**), source A (**ii**), and source B (**iii**) is given by the first row, and the predicted cochleagrams using the output connections of all hidden units (**iv**), of only G1 hidden units (**v**), and of only G2 hidden units (**vi**) is given by the second row. **A.** This triplet yielded the best SDR probe score of* 4.92*. For SDR, the higher metric denotes better separation; thus, the parallel pairing (G1-A, G2-B) – with* $M_1 = 4.92$ *– yields a better separation than the cross pairing (G1-B, G2-A) – with* $M_2 = 1.26$*. Qualitatively, the parallel pairing is noticeably better, as the low-frequency activity in sound A (**A.ii**) is predicted by G1 hidden units (**A.v**), and the high-frequency activity in sound B (**A.iii**) similarly appears in G2's prediction (**A.vi**). **B.** The triplet with the second best SDR probe score of 3.91, where the cross pairing yields a better separation (*$M_1 = 0.88, M_2 = 3.91$*). Qualitatively, the better fit of the cross pairing is difficult to see as both G1 (**B.v**) and G2 (**B.vi**) predictions possess the mid-frequency activity of sound A (**B.ii**) and sound B (**B.iii**) around 2s and 1s respectively. Yet, there is a small amount of low-frequency activity predicted by G1 (**B.v**) that corresponds with that of sound B (**B.iii**). **C.** The triplet with the third best SDR probe score of 3.83, where the cross pairing yields a slightly better separation (*$M_1 = 2.69, M_2 = 3.83$*). The high-frequency activity in sound A (**C.ii**) around 1.5s is more strongly predicted by G2 (**C.vi**), yet both G 1(**C.v**) and G2 (**C.vi**) predict the 1.6 kHz activity in sound B (**C.iii**) around 500s.*

Exploring the Qualities of Regularized Hidden Unit Populations

To better understand how distinctly regularized hidden unit populations give rise to decent, unsupervised source separation, a series of analysis was conducted that showed that G1 and G2 were tuned to distinct volume and frequency ranges.

Library and Party Hidden Units: Proclivity to Representing Sounds of Differing Volume

To explore whether one of the hidden unit populations (i.e. G1 and G2) reliably modeled the quieter sound while the other predicted the louder one, the mean amplitudes of G1 and G2's predictions for all validation mixture sounds were calculated. Figure 13 shows that on average, G1 predicted a softer sound with a lower mean amplitude, while G2 predicted a louder one. Additionally, when calculating SDR scores for each validation triplet example, 84.03% of the time, the better sound separation pairing occurred when G1 was paired with the single-speaker target sound with the lower mean predicted amplitude.

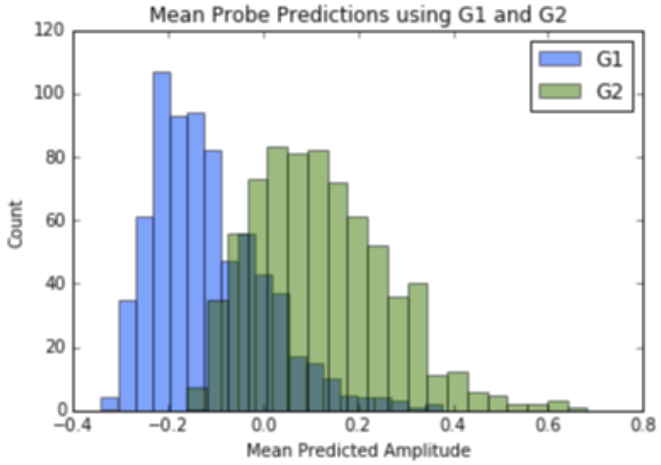


Figure 13: Amplitude Difference between G1 and G2's Probe Predictions

For every mixture in the validation set, the mean amplitude predicted by using just G1's output connections $(\mu = -0.12, \sigma = 0.12)$ *as well as that predicted by just using G2's were calculated* $(\mu =$

$0.12, \sigma = 0.14$). *The difference between the G1 and G2 distributions of mean predicted amplitude suggests that G1 predicts the softer sound in a mixture, while G2 predicts the louder one.*

How low can you go? Frequency Tuning of Different Hidden Unit Populations

Analysis of the mean amplitude of the predictions using G1 and G2 for each frequency band revealed more complexity. For each of the 39 log-spaced frequency bins ranging from 100 to 8093 Hz, the mean amplitudes for all validation mixture sounds predicted by G1 and G2 were calculated (Figure 14). The overall, mean amplitudes predicted by G1 and G2 from all validation mixture cochleagrams at each frequency band (Figure 14b) are strongly, inversely related ($r = -0.86, p < 10^{-6}$; see Figure 14 for more details on statistical testing), suggesting that they predict distinct activity at each frequency band instead of predicting equal amounts of activity at each frequency band. Further analysis shows that both G1's and G2's overall mean predictions per-frequency are significantly correlated in opposite directions ($r = -0.36, p = 0.02$ for G1; $r = 0.48, p = 0.002$ for G2) to the values of the frequency bands used. These opposite-direction correlations suggest that G2 may predict more high frequency activity, while G1 predicts more low frequency activity.

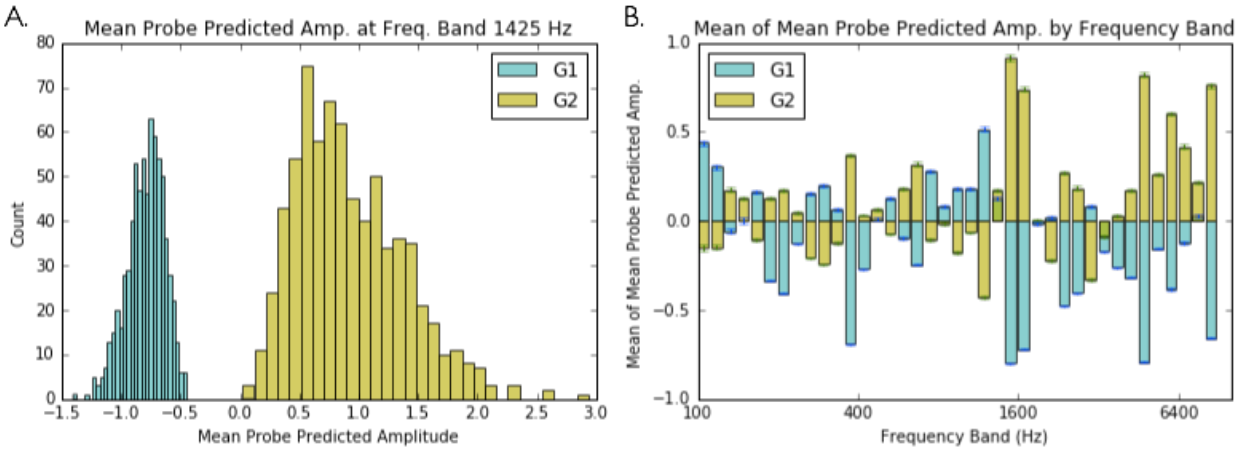


Figure 14: Frequency Differences between G1 and G2 Predictions and between their Best Matched Targets

D. *For all mixtures in the validation examples, the mean amplitude at the 1425 Hz frequency band predicted by G1 ($\mu = -0.80, \sigma = 0.16$) and G2 ($\mu = 0.92, \sigma = 0.45$) was binned into a histogram. The difference in the 1425 Hz mean amplitude predictions of G1 and G2 suggests that G2 predicts more activity at 1425 Hz while G1 predicts negative, activity.* ***B.*** *The mean amplitude at each frequency band predicted by G1 and G2 was calculated for all validation mixtures, and the mean and standard error of those mean amplitudes at each frequency was plotted. For example, from* ***A****, the mean of the mean amplitudes predicted by G1 at 1425 Hz is* -0.80*, while that of G2 is 0.92. While it is hard to discern a clear pattern between frequency and G1 or G2's mean of mean predicted amplitudes, for frequencies greater than 1300 Hz, G2 appears to predict more positive activity than G1.* ***Statistical Testing.*** *Yet, inference testing of computed correlation coefficients suggests significant relationships: the correlation coefficient between G1's mean of mean amplitudes per frequency band and G2's is* -0.86 *($p < 10^{-6}$), the correlation coefficient between the frequency bands and G1's per-frequency mean of mean amplitudes is* -0.36 *($p = 0.02$), and the correlation coefficient between the frequency bands and G2's is* 0.48 *($p = 0.002$).*

However, Figure 14b suggests a more complex situation, with the opposite, low-high frequency tuning of G1 and G2 appearing less strong in the lower frequency bands between

100-800Hz. Coupled with the facts that first, G1 and G2's overall mean predictions are reliably inversely related, and second, cochleagram activity is not evenly distributed across frequency bands (Supplementary Figure 24), this observation invites further research in on how the mammalian auditory system separates out sounds with respect to frequency as well as how well the frequency statistics of human speech and other natural sounds match up with the unsupervised source separated predictions.

A Tale of Many Relationships: Predictive Ability vs. Source Segregation Capability vs. Hidden Unit Population Regularization Factors

Until now, the analysis described in this dissertation focused on one fully-connected model with distinctly regularized hidden unit populations. Analyzing the separate hyper-parameter searches for α and β^*, where $\beta = \beta^*\alpha$ and α and β are the regularization factors on G1 and G2, reveals interesting relationships among supervised predictive quality, unsupervised source separation ability, and inter- and intra- hidden unit regularization.

Figure 15a,c,e shows that, as α increases from 10^{-10} to 10^{-3} while $\beta^* = 1.2$, a fully connected model with hidden unit regularization improves in its source separation while simultaneously worsening in its predictive ability and that predictive power and source segregation quality is significantly inversely correlated as α varies ($p = 0.04$). However, as β^* varies from 2^0 to 2^9 while $\alpha = 10^{-7}$, Figure 15b,d,f suggests that an optimum β^* can be found and that predictive ability and sound separation capacity are strongly, positively correlated as β^* varies ($p = 0.02$).

Intuitively, as α increases, the amount of regularization on both hidden unit populations increase and the $\beta^* = 1.2$ factor difference in regularization appears to be enough to improve source separation quality when α is strong enough. Additionally, it makes sense that more hidden unit regularization with increased α would restrict the hidden units' ability to predict the future because their activity is being penalized.

However, the clear optimal β^* and positive relationship between predictive quality and source segregation ability in the β^*-varying search is more challenging to unify. Recall that β^* describes relatively how much G1 and G2 are regularized differently. With a relatively small $\alpha = 10^{-7}$, it makes sense that small β^* values yield poor source separation because the combination of weak overall regularization by α and weak differentiated regularization by β^* is insufficient to produce distinct predictions. It could be plausible that the target single-speaker sounds are similar enough that a large β^* drives the predictions too far apart and thus worsens source separation capacity because one prediction is too far off from its target. The positive relationship between prediction and source segregation quality could be explained by the idea that, when the right β^* is found to appropriately differentiate G1 and G2, not only does sound separation improve but predictive ability does as well because an improved ability to represent different sound sources leads to improved predictions. However, further research on the effects of hidden unit regularization factors, including 2D and finer-grain α and β^* grid searches, is needed to test these hypotheses.

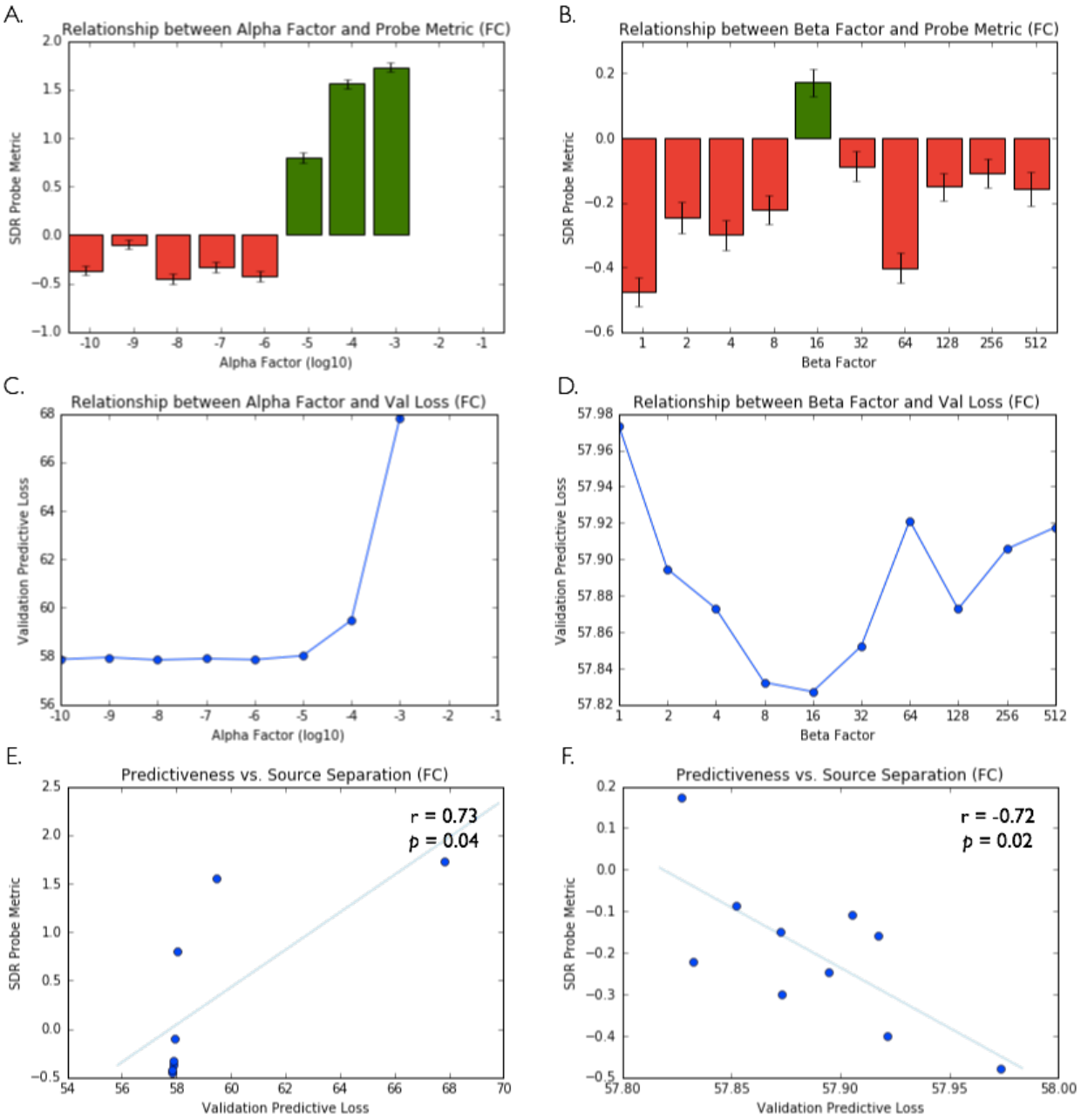


Figure 15: Relating Source Separation, "Predictive"-ness, and Hidden Unit Regularization Factors (FC)

Left Column (A,C,E). *10 FC models with explicit hidden unit populations were trained on the predictive task with the same parameters as the one used for in-depth analysis, except that* $\alpha \in \{10^{-10}, \dots, 10^{-1}\}$ *and* $\beta = 1.2\alpha$*. Two models (*$\alpha = 10^{-2}$ *and* $\alpha = 10^{-1}$*) yielded invalid NaN (not a number) predictions after training and were excluded.* ***A.*** *The log value of each model's α value was plotted against its SDR probe metric – the mean SDR score over all validation triplet examples and a measure quantifies how well a model performs on unsupervised source separation. A positive SDR score, which models with* $\alpha \in \{10^{-5}, 10^{-4}, 10^{-3}\}$ *yielded, suggests that more signal than noise is captured in the optimally predicted source separation.* ***C.*** *The log value of each model's α value was plotted against its validation loss on the supervised predictive task. The predictive loss palpably worsens for models with* $\alpha \in \{10^{-4}, 10^{-3}\}$*. Coupled with observations from* ***A****, there appears to be a negative relationship between predictive power and source separation ability as α varies.* ***E.*** *Predictive loss was plotted against SDR score for the 8 well-defined, α-varying FC models; the computed correlation coefficient and a two-sided t-test testing the null hypothesis that the fitted line has slope = 0 suggests that, as α varies, "predictive"-ness worsens (i.e.*

predictive validation loss increases) as source separation improves (i.e. SDR increases). ***Right Column (B,D,F).*** *10 FC models with explicit hidden unit populations were trained on the predictive task with the default FC hyper-parameters (with $\alpha = 10^{-7}$), while β varies, $\beta \in \{2^0\alpha, \dots, 2^9\alpha\}$. For simplicity, let beta factor $\beta^* \in \{2^0, \dots, 2^9\}$ refer to the scaling factor on α.* ***B.*** *Each model's β^* value was plotted against its SDR probe metric; with only the $\beta^* = 16$ model, which is the model that was analyzed in-depth, yielding a positive SDR score.* ***D.*** *Each model's β^* value was plotted against its predictive validation loss, with the best loss occurring at $\beta^* = 16$. Coupled with observations from* ***A****, there appears to be a positive relationship between predictive power and source separation ability as β^* varies.* ***F.*** *Predictive loss was plotted against SDR score for the 10 β^*-varying FC models; statistical testing (same as that for* ***E****) suggests that "predictive"-ness and source separation are positively linked as β^* varies.*

Can You Hear Me Crystal Clear? Room for Improvement in Unsupervised Sound Source Separation

Further analysis on the FC network with $\alpha = 10^{-7}$ and $\beta^* = 16$ shows that there is still ample room for improving its unsupervised source separation. Recall that a negative SDR measure suggests that there is more error than signal captured in the source separation. 49.4% of the 720 validation triplets yielded negative overall SDR scores, implying that nearly half the separated predictions contained more error than signal (Supplementary Figure 25). Depicting the three examples with the top three worst SDR scores, Figure 16 shows that either source separation is hard to observe by the naked eye or the optimal pairing between G1 and G2 predictions to sound A and B targets given by the SDR score seems counter-intuitive to a human viewer (Figure 16b).

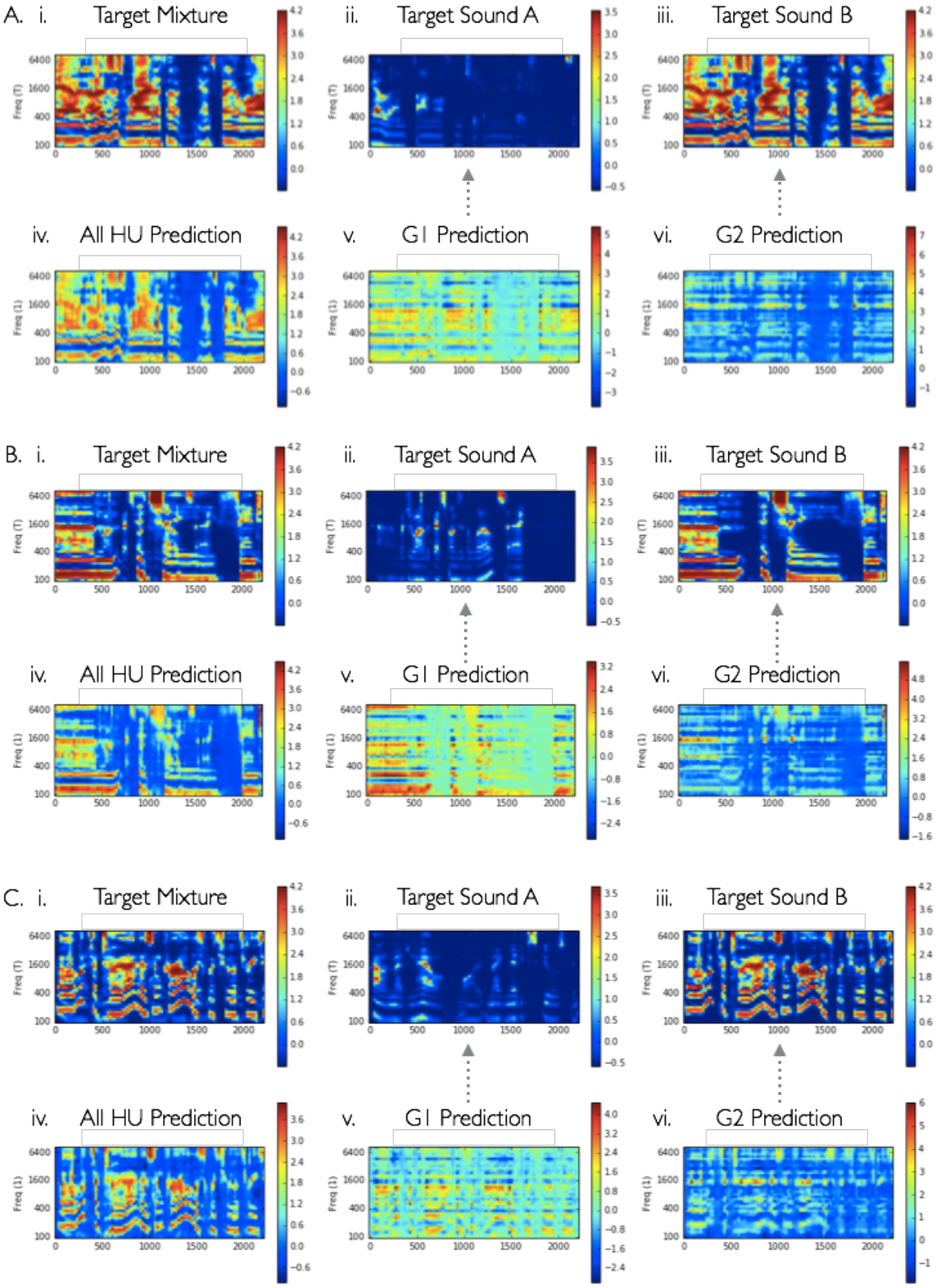

A. i. Target Mixture
ii. Target Sound A
iii. Target Sound B
iv. All HU Prediction
v. G1 Prediction
vi. G2 Prediction
B. i. Target Mixture
ii. Target Sound A
iii. Target Sound B
iv. All HU Prediction
v. G1 Prediction
vi. G2 Prediction
C. i. Target Mixture
ii. Target Sound A
iii. Target Sound B
iv. All HU Prediction
v. G1 Prediction
vi. G2 Prediction
Freq (T)
Freq (1)

Figure 16: Examples of Unsupervised Source Separation with the Worst SDR Probe Measures (SDR)

*For each example (**A, B, C**), the target cochleagrams for the mixture (**i**), source A (**ii**), and source B (**iii**) is given by the first row, and the predicted cochleagrams using the output connections of all hidden units (**iv**), of only G1 hidden units (**v**), and of only G2 hidden units (**vi**) is given by the second row. **A.** This triplet yielded the worst SDR probe score of -2.83 ($M_1 = -2.83, M_2 = -6.72$), with the parallel pairing (G1-A, G2-B) producing the better source separation output because $M_1 > M_2$; however, it is qualitatively difficult to determine which pairing produces the best source separation, as both G1 (**A.v**) and G2 (**A.vi**) appear to predict parts of sound B's activity (**A.iii**). **B.** The triplet with the second worst SDR probe score of -2.41 ($M_1 = -2.41, M_2 = -3.88$), where the parallel pairing yielded the better separation. Qualitatively, the cross pairing actually looks better, as G1 (**B.v**) appears to predict the spanning frequency activity in the first 600 ms of sound B (**B.iii**). However, using the SDR measure, G1 barely predicts sound B better than G2 ($SDR(G1, B) = 1.49$, $SDR(G2, B) = 1.48$). **C.** The triplet with the third worst SDR probe score of -2.37 ($M_1 = -2.37, M_2 = -5.86$), where the parallel pairing yielded the better source separation. Yet, both G1 (**C.v**) and G2 (**C.vi**) appear to faintly predict frequency-spanning activity in sound B (**C.iii**).*

Performance using RNNs

Like the $\alpha = 10^{-7}, \beta^* = 16$ FC model, a RNN model with the following hyper-parameters was found to have a competitive predictive loss from a series of grid searches and chosen to be analyzed in-depth for its sound source capabilities:

- Update function: adam
- Non-linearity: tanh
- L1 regularization factor on network weights: 10^{-4}
- Regularization factors for explicit hidden unit populations (G1, G2, H): $\alpha = 10^{-7}, \beta = 2\alpha, \gamma = \frac{\alpha\sqrt{\frac{|G2|}{|H|}U}}{2U}$

However, unlike the FC model analyzed in-depth, the RNN's overall SDR score of -0.38 suggests that the unsupervised source separation by its regularized G1 and G2 populations contained more error than signal. Furthermore, Supplementary Figure 26a explains how its SDR score compared poorly to scores generated by 10,000 random splits of two hidden unit groups in a permutation test ($p = 0.50$), while Supplementary Figure 26b shows that only 27.3% of the 720 validation triplets earned positive SDR scores when using the RNN for source separation. In comparison, the FC model had an overall SDR score of 0.17, with 50.6% of examples earning positive scores, and performed significantly better than random hidden unit groups ($p = 0.035$).

Figure 17 shows the relationships between predictive ability, source separation capacity, and the hidden unit regularization factors α and β^*; Figure 17a,c suggests a similar trend as that observed in the α-varying grid search for FC models (Figure 15) that, as α increase, predictive power worsens while source separation capacity improves, until very large $\alpha = 10^{-1}$, when it worsens again. However, unlike the $p < 0.05$ relationships identified in the α and β^*-varying FC grid searches, Figure 17e,f show that the RNN trends fail to survive significance testing. Furthermore, the lack of a clear trend between β^* and the SDR score (Figure 17b) suggests that the $\alpha = 10^{-7}$ held constant in the β^*-varying grid search was not a suitable parameter for the RNN. One would expect RNNs to perform better than FC

models on tasks involving temporal data; however, they are known to be harder to initialize (Le et al., 2015), so in addition to performing more dense hyper-parameter grid searches for well-paired α and β^*, further work should be done to initialize the RNN well.

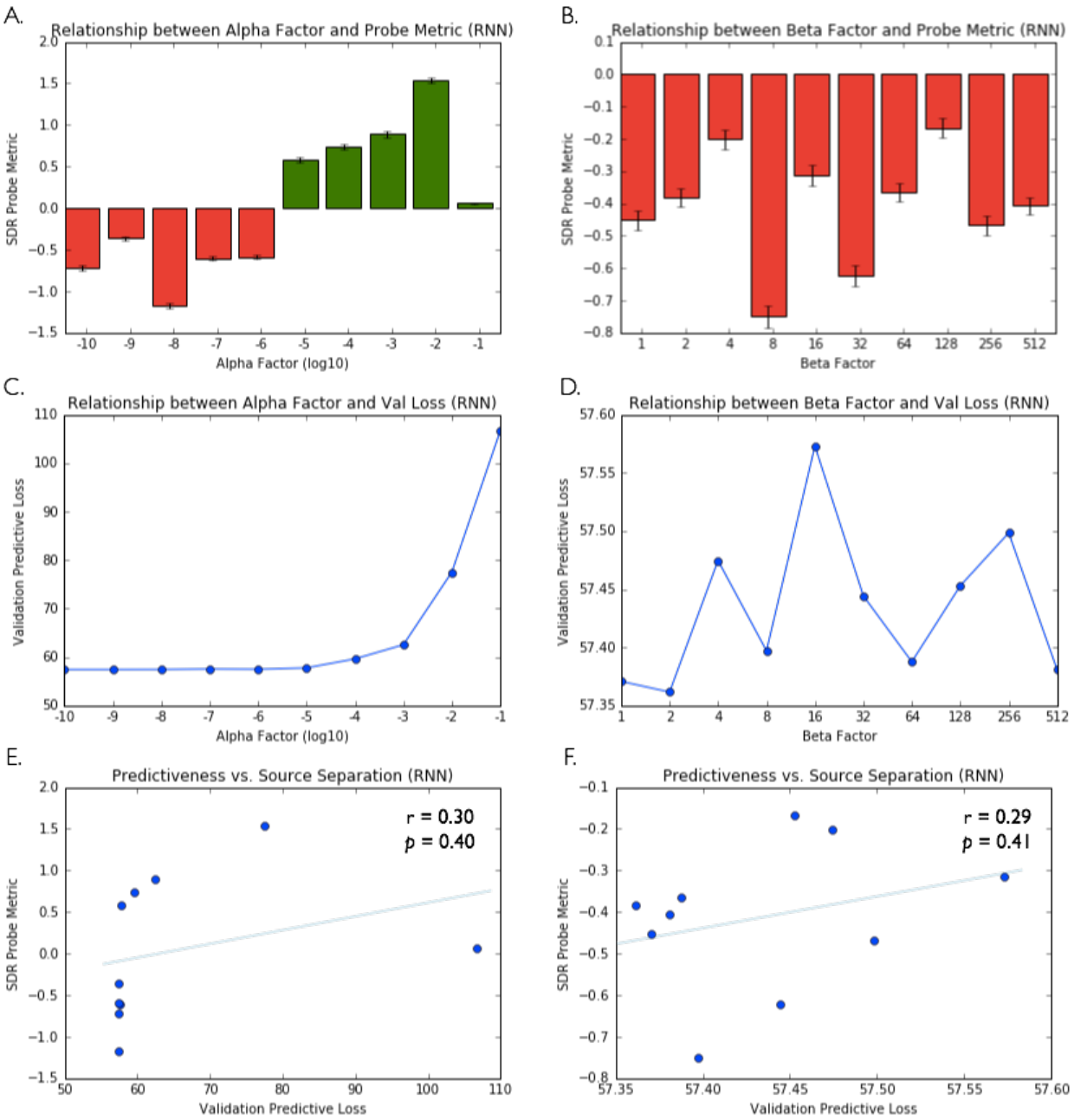


Figure 17: Relating Source Separation, "Predictive"-ness, and Hidden Unit Regularization Factors (RNN)

Left Column (A,C,E). *10 RNN models with explicit hidden unit populations were trained on the predictive task with the same hyper-parameters as the one used for in-depth analysis, except that* $\alpha \in \{10^{-10}, \ldots, 10^{-1}\}$ *and* $\beta = 1.2\alpha$. ***A.*** *The log value of each model's α value was plotted against its SDR probe metric – the mean SDR score over all validation triplet examples and a measure quantifies how well a model performs on unsupervised source separation. A positive SDR score, which models with* $\alpha \geq 10^{-5}$ *yielded (although the* $\alpha = 10^{-1}$ *model's score barely exceeded zero), suggests that more signal than noise*

is captured in the optimally predicted source separation. ***C.*** *The log value of each model's α value was plotted against its validation loss on the supervised predictive task. The predictive loss palpably worsens for models with* $\alpha \geq 10^{-4}$*. Coupled with observations from* ***A****, there appears to be a negative relationship between predictive power and source separation ability as α varies.* ***E.*** *Predictive loss was plotted against SDR score for the 10, α-varying RNN models; the computed correlation coefficient and a corresponding two-sided t-test testing the null hypothesis that the fitted line has slope = 0 were not statistically significant to affirm the hypothesis that, as α varies in RNN models, "predictive"-ness worsens (i.e. predictive validation loss increases) as source separation improves (i.e. SDR increases).* ***Right Column (B,D,F).*** *10 RNN models with explicit hidden unit populations were trained on the predictive task with the default RNN hyper-parameters (with* $\alpha = 10^{-7}$*), while β varies,* $\beta \in \{2^0\alpha, \dots, 2^9\alpha\}$*. For simplicity, let beta factor* $\beta^* \in \{2^0, \dots, 2^9\}$ *refer to the scaling factor on α.* ***B.*** *Each model's* β^* *value was plotted against its SDR probe metric, with no easily observable relationship between* β^**and source separation.* ***D.*** *Each model's* β^* *value was plotted against its predictive validation loss, with the best loss occurring at* $\beta^* = 2$ *(the* $\beta^* = 2$ *RNN model was used for in-depth analysis).* ***F.*** *Predictive loss was plotted against SDR score for the 10* β^**-varying RNN models; statistical testing (same as that for* ***E****) did not show any significant relationship between "predictive"-ness and source separation as* β^* *varies in these RNN models. These set of sub-figures can be compared to those in Figure 15 respectively, which contain corresponding figures using α and* β^**-varying FC models.*

FC and RNN Models with the Best Sound Separation but Poor Predictive Ability

Figure 15 and Figure 17 identified FC and RNN models respectively with better sound separation (but poorer predictive power) than the models analyzed in-depth. Good sound separation examples from the models with the best sound separation – the $\alpha = 10^{-3}$ FC and $\alpha = 10^{-2}$ RNN models with $\beta^* = 1.2$ for both – are shown in Figure 18. Compared to the examples for the $\alpha = 10^{-7}, \beta^* = 16$ FC model (Figure 12), the predictions made by G1 and G2 are more distinct from one other, which makes sense because a larger α increases the amount of regularization on and competition between the two populations. Nevertheless, the predictive qualities of superior sound-separating networks are clearly worse, as the predictions are blurrier and coarser. Comparing between these two superior sound-separating FC and RNN networks, the fully-connected model outperforms the recurrent one both in its predictive ability, with a validation loss of 67.82 compared to the RNN's loss of 77.50, and in its source separation capacity, with a better overall SDR score of 1.73 compared to the RNN's score of 1.54. This surprising outperformance suggests that further work can be done to improve the RNN in its initialization and hyper-parameter settings. Nevertheless, the strength of these superior, source separating models points to the promise of this paradigm for using hidden units to perform unsupervised sound source separation as well as of the network extension of distinctly regularization the activity of certain hidden units.

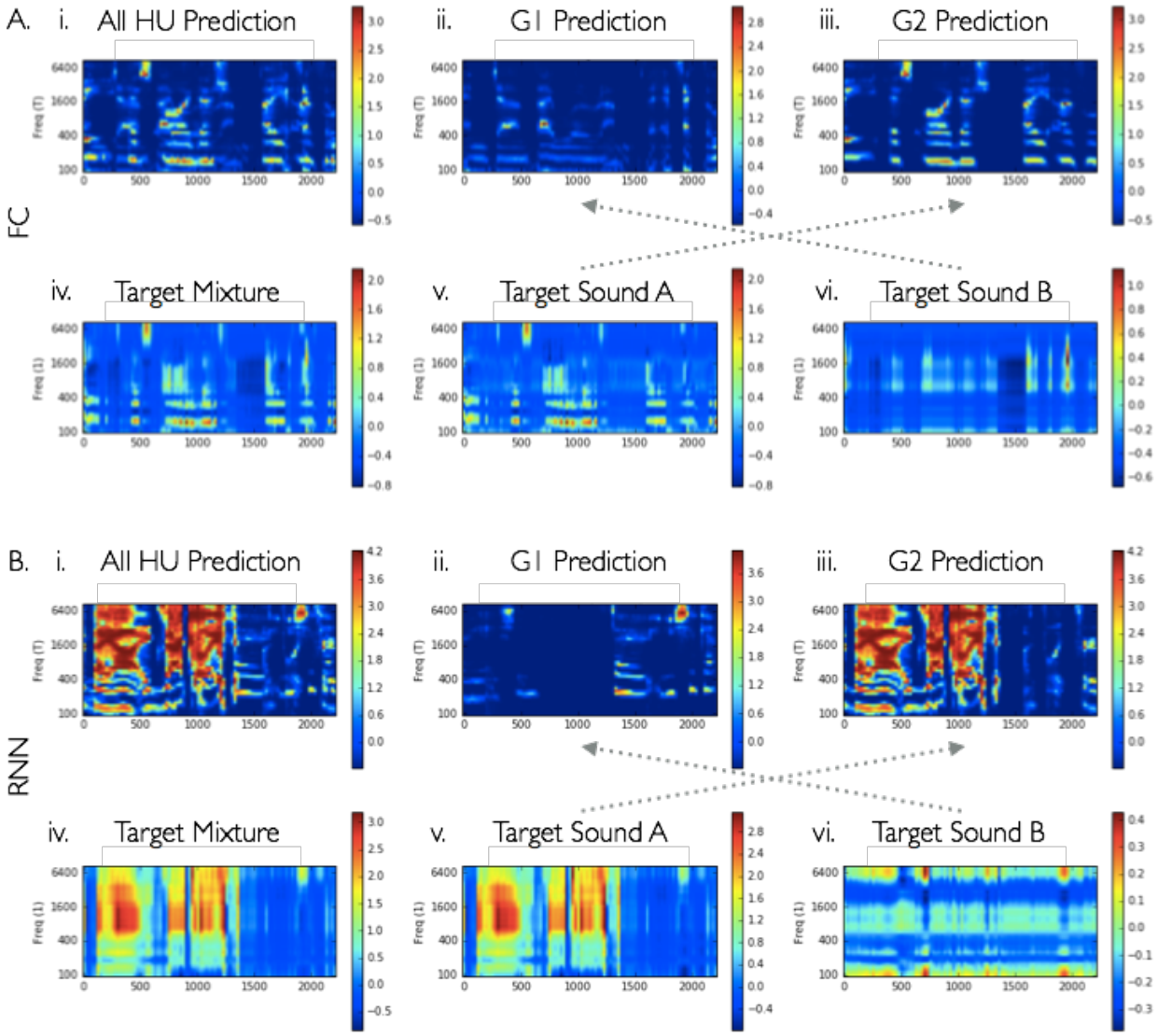


Figure 18: Source Separation Examples for the FC and RNN Models from the α-varying Grid Search with the Best Mean SDR Probe Score

*For each example (**A, B**), the target cochleagrams for the mixture (**i**), source A (**ii**), and source B (**iii**) is given by the first row, and the predicted cochleagrams using the output connections of all hidden units (**iv**), of only G1 hidden units (**v**), and of only G2 hidden units (**vi**) is given by the second row. **A.** Identified in an α-varying grid search described in Figure 15 as the FC model with the best SDR score of* 1.73*, a FC in-depth-analysis model with $\alpha = 10^{-3}, \beta = 1.2\alpha$, and otherwise default FC hyper-parameters was used to probe the validation set for its source separation performance. This triplet yielded the best SDR probe score of* 7.08 *($M_1 = 3.22, M_2 = 7.08$), with the cross pairing yielding the better source separation because $M_2 > M_1$. Qualitatively, the cross pairing is noticeably better, as the low-frequency activity in sound B (**A.iii**) is predicted by G1 hidden units (**A.v**), whereas the frequency-spanning activity near 2s in sound A (**A.ii**) similarly appears in G2's prediction (**A.vi**). **B.** Identified in an α-varying grid search described in Figure 17 as the RNN in-depth-analysis model with the best SDR score of* 1.54*, a RNN model with $\alpha = 10^{-2}, \beta = 1.2\alpha$, and otherwise default RNN hyper-parameters was used to probe the validation set for its source separation performance. This triplet yielded the fourth[3] best SDR probe score of* 3.75 *($M_1 = -7.04, M_2 = 3.75$), where the cross pairing yielded the better separation. Qualitatively, the cross pairing*

[3] The fourth best SDR probe score example was chosen over those corresponding to the first, second, or third for illustrative purposes, the three other better examples did not have many defining amplitude and frequency characteristics.

*displays some clear source separation characteristics: G1 (**B.v**) predicts well the early broadband activity in the first half of sound B (**B.iii**). While G2 predicts some extraneous activity along with the high frequency activity right before 2s in sound A, it is notable that the G1 (**v**) and G2 (**vi**) predictions in both **A** and **B** are more distinct from each other than those generated by the in-depth-analysis FC model (Figure 12).*

Biologically-Consistent Spectro-Temporal Receptive Fields (STRFs) of Hidden Units

Without constraints enforcing biological characteristics, the hidden units of networks trained on the predictive tasks exhibited realistic receptive fields and included features like lagging and flanking inhibition (Figure 19). Singer et al., n.d. first showed that naturalistic receptive fields arose in fully-connected networks trained on both visual and auditory predictive tasks; this dissertation not only confirms that result in FC models but also demonstrates that predictive RNNs also exhibit the same biologically-consistent qualities (Figure 19c). Figure 19 includes receptive field visualizations that are derived by reshaping the input and output weight matrices of neural networks; however, a more accurate visualization of RNN receptive fields would incorporate its recurrent connections by visualizing the best linear mapping from cochleagram to hidden unit activity. These results bolster the idea that mammalian receptive fields are tuned to qualities most important for predicting the future.

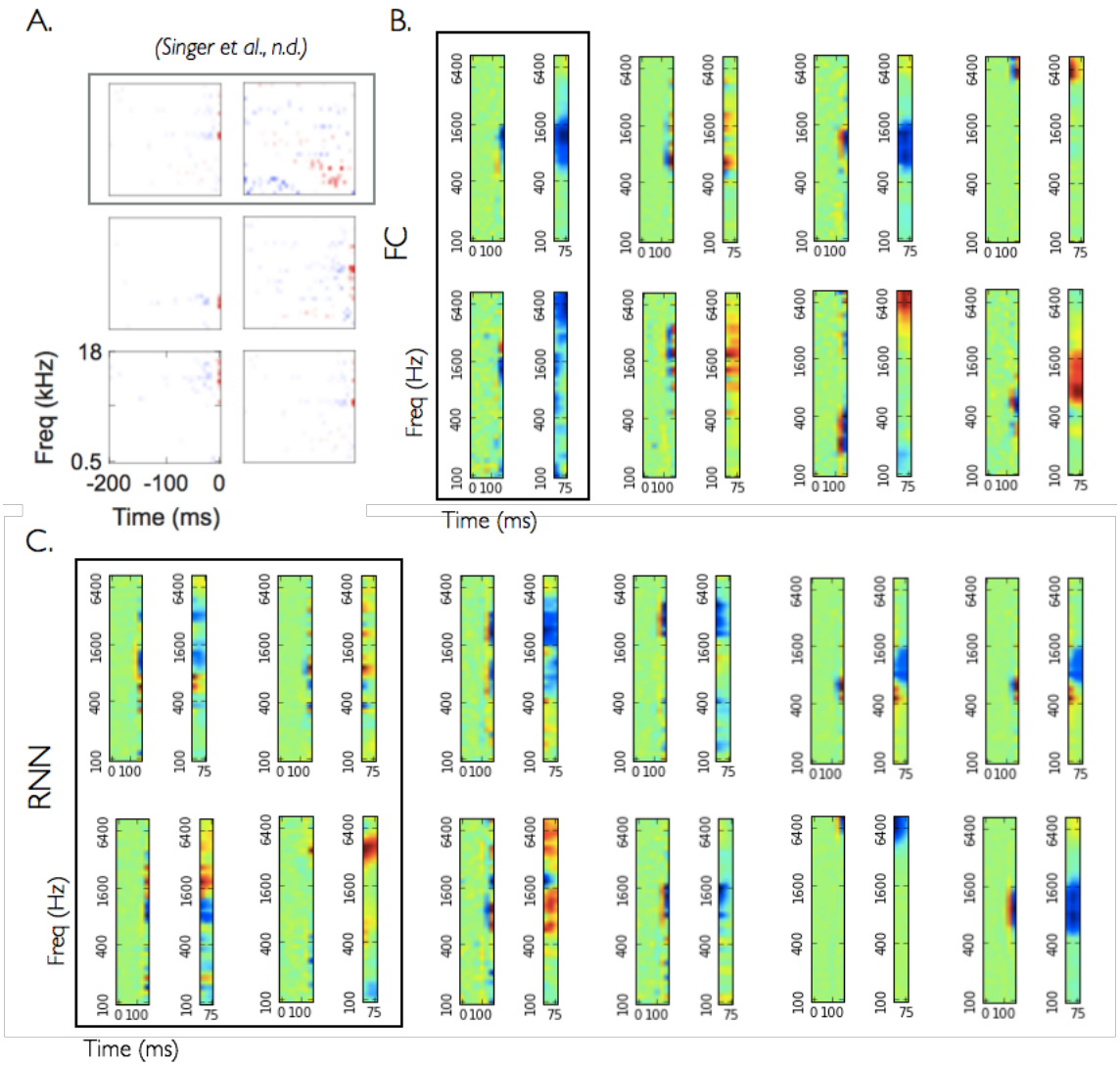


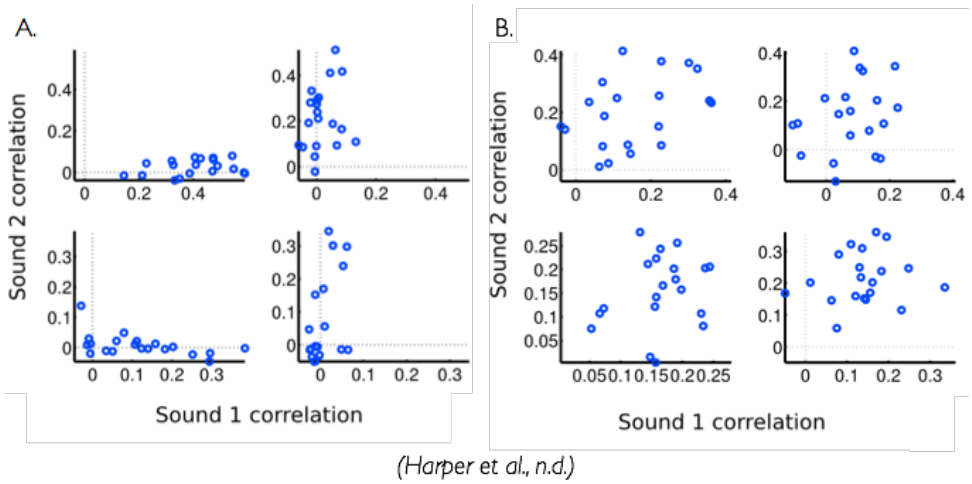
Figure 19: STRFs of Real Ferret Neurons and FC and RNN Hidden Units

A. *STRFs from recorded auditory cortex neurons in ferrets* (Singer et al., n.d.)*. Two examples of flanking inhibition, where there is excitatory and inhibitory activity in nearby, flanking frequency bands, is boxed in grey. The other four examples demonstrate lagging excitation, where exciting activity quickly follows inhibitory activity.* ***B and C.*** *Visualizations of the fan-in and fan-out activity of 6 hidden units from the in-depth-analysis FC model (**B**) and 12 hidden units from the in-depth-analysis RNN model (**C**) were selected for their biologically-consistent qualities. In both **B** and **C**, hidden units with flanking inhibition are boxed in black. Non-boxed examples exhibit lagging inhibition or excitation, wherein excitation or inhibition is quickly followed by inhibition or excitation respectively. Note that activity is spatially and temporally tuned, as in real neurons (**A**).*

Biologically-Consistent, Selective Hidden Units

In addition to having realistic receptive fields, the hidden units of the predictive networks with distinctly regularized hidden unit populations also showed a range of selectivity similar to that observed in primary auditory cortex neurons of ferrets. Figure 20a demonstrates how a real, "selective" neuron reliably responds to a mixture and its component sounds: its responses to one component sound are highly correlated to those to

its mixture sound but its responses to the other component sound are not correlated to those of its mixture sound. Observations of such neurons from Harper et al., n.d. suggest that some primary auditory cortex neurons play a role in sound selection by reliably responding to one single sound but not the other in a similar fashion to how they respond to the mixture of the two sounds. Figure 20b shows that a "non-selective" neuron's responses to two individual component sounds are similarly correlated to its responses to the mixture sound.



(Harper et al., n.d.)

Figure 20: High and Low Selective Auditory Cortex Neurons from Anesthetized Ferrets

*The set of plots in **A** and **B** each corresponds to one neuron recorded in the auditory cortex of an anesthetized ferret. Each plot represents a distinct triplet set of sounds (sound 1, sound 2, and mixture of sound 1 and 2), each of which was played to the ferret 20 times. For each repeat, the correlation coefficient was calculated between the peristimulus time histogram (PSTH) (Figure 11d) of the neuron when sound 1 is played and the mean PSTH of the neuron when the mixture is played. Correlation coefficients between PSTHs corresponding to 20 repeats of sound 2 and the mean mixture PSTH were also calculated, and the correlations were plotted against each other. The neuron in **A** demonstrates that it's selective to a single sound because for a given triplet of sounds, it reliably shows no correlation between one sound and the mixture yet consistent correlation between the other sound and the mixture; that is, its responses in a similar way to one sound as it does to the mixture but not in a correlated way for the other sound. In contrast, the neuron in **B** exhibits low selectivity because there are few instances where it selectively responds to one sound in a similar way to the mixture but not the other sound.*

Figure 21 shows that both FC and RNN predictive networks with hidden unit regularization exhibits similar patterns of low and high selectivity as real neurons. The examples of high selectivity in the hidden units expressed selectivity differently from those in real neurons in that strong correlations between a single sound and its mixture were more often observed (i.e. points near (1,0) and (0,1) in Figure 21's visualizations) compared to no correlation between one sound and the mixture (i.e. points along the x- and y- axes in Figure 20's plots). Similar patterns of high and low selectivity were also found in typically-defined FC and RNN networks, suggesting that an efficient encoding of predictive qualities of sound may also encode some representation of separated sounds.

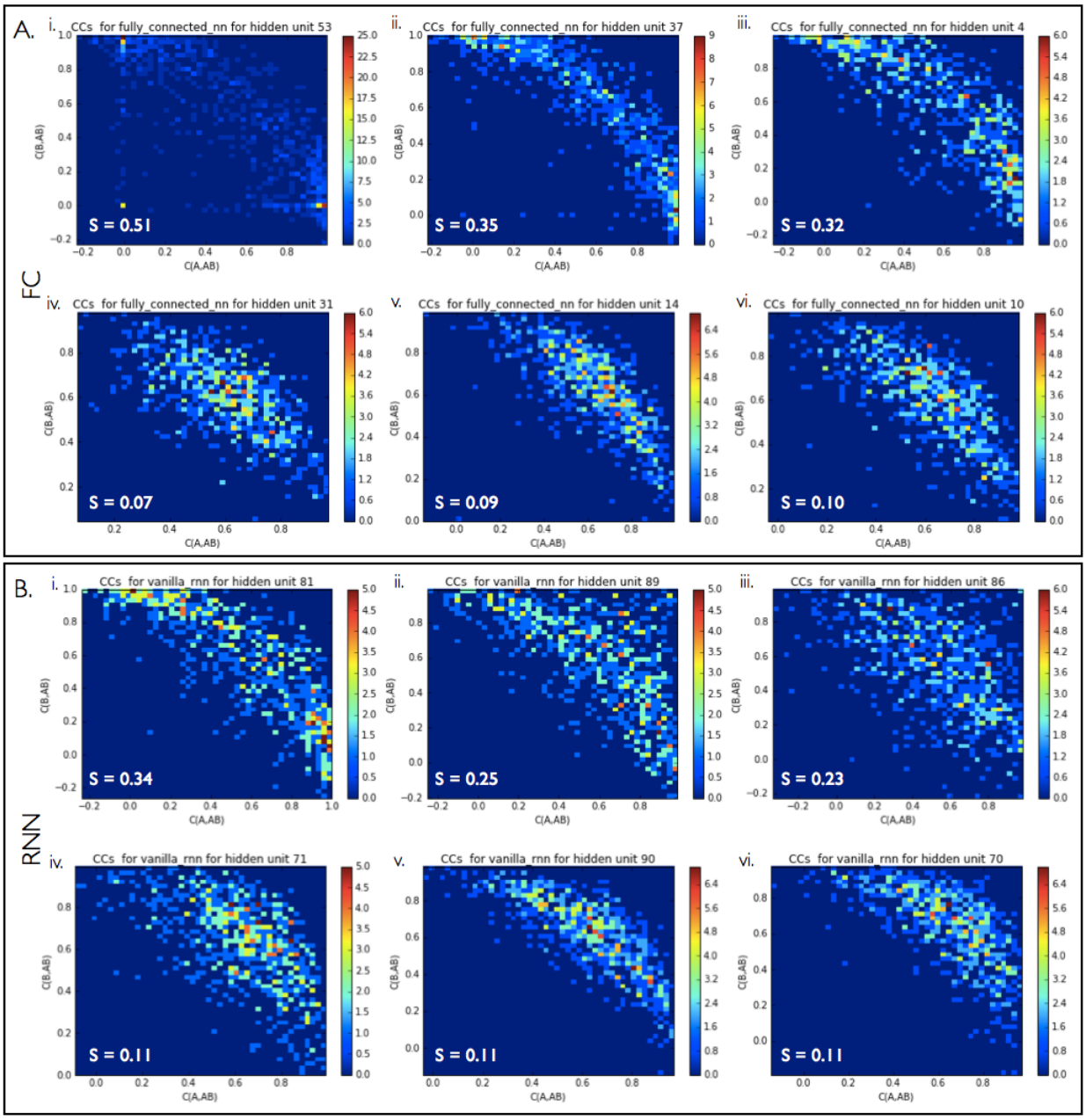


Figure 21: Top 3 High and Low Selective Hidden Units in FC and RNN models

*The three hidden units with the highest aggregate selectivity scores in the in-depth-analysis FC (**A**) and RNN (**B**) model are visualized in the top rows (**i,ii,iii**) of their respective panels, while the three hidden units with the lowest aggregate selectivity scores are visualized in the panel's bottom rows (**iv,v,vi**). For every validation triplet example, the correlation between sound A and the mixture AB is plotted against that between sound B and the mixture AB. Having many coordinates fall near (1,0) or (0,1) suggests that the hidden unit is selective for a single sound because it is highly correlated with one sound and not correlated with the other sound in a given example. Particularly in the FC model, the most selective hidden units show similar qualities to the selective AC neurons in ferrets (Figure 20) of having many points fall along the x- or y- axis.*

In the FC and RNN models with the best source separation scores from the α-varying grid search, strong source separation was linked to different qualities of the network. In the $\alpha = 10^{-3}, \beta = 1.2$ FC model, only a little more than 10 of the 80 hidden units with output connections were being used: the other units had near-zero selectivity and random-looking receptive fields. However, of the active units, at least 5 of them had strong selectivity qualities, with aggregate selectivity scores greater than 0.3. In contrast, in the RNN, no single or few hidden units stood out with strong selectivity; all units had aggregate selectivity scores between 0.10 and 0.19. These observations suggest that the mechanisms of sound separation in FC and RNN models may differ in how they use selective units; furthermore, the surprisingly phenomenon of a few hidden units contributing to most of the FC network's output suggests that strong selective units are important for strong source separation.

For the in-depth analysis FC and RNN models, again there are different distributions of selectivity. Supplementary Figure 27a shows that the FC model had more high selectivity units than the RNN model, which may have contributed to the FC model's superior sound separation ability. Supplementary Figure 27b,c show that the distribution of selective units between the FC and RNN models differed, with the RNN's most selective units coming from the H population that has no output connections, only recurrent ones to other hidden units. This observation suggests that the recurrent connections in the RNN's H population may be modulating the G1 and G2 populations and thereby using the H population's selectivity to improve sound separation. However, further research into the relationship between and mechanisms of selectivity and sound separation must be conducted in order to test these hypotheses.

Discussion (max. 2000 words)

This work demonstrates the promise of a novel paradigm for unsupervised sound source separation using distinctly regularized populations of hidden units in a predictive neural network. Furthermore, it shows that biologically-consistent receptive fields and selective characteristics can be identified in such network's hidden units.

Implications

These results have several implications on the existing literature on predictive coding, sound source separation, and neural networks. The promising, initial success of the presented unsupervised sound source separation paradigm that trains networks on a predictive task suggests that predictive coding may not only capture biologically-consistent characteristics such as realistic receptive fields (Singer et al., n.d.) and selective neurons (Harper et al., n.d.) but also encode information necessary not only for prediction but also sound separation, a biologically-necessary ability. This work represents the first time predictive coding was demonstrated to prove useful another essential task for mammalian survival.

Additionally, the presented paradigm is the first known generalizable unsupervised approach; the only other known unsupervised work used top-down, hand-tuned filters that relied on harmonic qualities in sounds, thereby limiting its application to tonal sounds, which excludes many natural sounds (Elhilali and Shamma, 2008). In contrast, our approach is far more unconstrained and thus can accommodate a greater, more realistic diversity of sounds.

Finally, as neural networks have gained prominence for their ability to perform well on well-defined, supervised tasks like image recognition, the machine learning research community is interested in exploring whether neural networks can be used to tackle more general problems like continual knowledge acquisition and representation that may be less defined or unsupervised (Mitchell et al., 2015). Regularizing different populations of hidden units may be a generalizable technique and be used to retool trained networks on supervised tasks to perform related unsupervised ones. For instance, a visual analog to the auditory problem of source separation is object segmentation in images; variants of this regularization technique may be able to localize the different objects in an image using a network trained video frame prediction.

Future Directions

There are several research directions to better understand the limits and mechanisms of this unsupervised source separation paradigm as well as to explore other network extensions besides or in addition to regularization hidden unit populations. Further computational work should also be done in close collaboration with similarly designed experimental research for a more direct comparison between the two realms. The experimental work compared to in this dissertation used different datasets as well as different time and frequency bins (Harper et al., n.d.; Singer et al., n.d.).

Optimizing Performance

Due to time constraints, sequential 1D hyper-parameter searches were used to set α and β^* parameters for regularizing the G1 and G2 hidden unit populations. To both better understand the relationships α and β^* have on predictive and source separation quality as well as to find the best regularization parameters that perform well on both the unsupervised and supervised problem, a 2D grid search for α and β^* should be conducted.

Furthermore, it was surprising that the RNN models, which have short-term memory capacity, did not perform better than the FC ones. It is known that RNNs can be difficult to initialize; thus, further work should be done to identify the best parameters and initialization scheme for RNNs to perform better than FC models on both the supervised predictive task as well as the unsupervised source separation task. One such initialization technique involves initializing the network's recurrent weights matrix to be a scaled identity matrix (Le et al., 2015), which proved promising for improving typically defined RNNs but was not tested for RNNs with distinctly regularized hidden unit populations. Another RNN model that allows for more stable training and longer memory capacity is the Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997). Inspired by human working memory, it has become the main RNN model used in machine learning research. Initial explorations for this project included LSTMs but were not continued due to limited computing power, yet future work should include LSTM models.

Varying the Dataset and Model Extension

Another avenue to explore includes varying the dataset. First, to compare our unsupervised paradigm to current state-of-the-art results, the classic sound separation datasets for speech separation between a male and female speaker (Kabal, 2002), singing separation between a singer and background music (Hsu and Jang, 2010), and speech denoising between a human speaker and background noise (Victor et al., 1990) should be used. In a few respects, our dataset and set-up were more difficult than the aforementioned problems: for instance, our speech separation dataset included both two-speaker mixtures without regard to gender as well as one-speaker sounds. The robustness of the model can also be tested by observing how much results deteriorate when increasing amounts of white noise, which would be hard to predict, is added to the dataset.

The limits of the paradigm can be tested by varying the ratio of mixtures and single sources in the training set as well as by increasing the number of single sources in a mixture and observing whether G1 and G2 separate out the two loudest sounds. The generalizability of this paradigm in this respect can also be tested by varying the number of and ratio of hidden units among hidden unit populations (i.e. G1, G2, G3, …) while increasing and varying the number of sources in mixtures in the training dataset.

Finally, more complex networks with multiple layers – and incorporating distinctly regularized hidden unit populations into them – should be explored. Some evidence suggests that source selectivity arises and increase more downstream in the auditory

cortex (Zion Golumbic et al., 2013). Thus, in addition to exploring whether additional layers increasing unsupervised source separation performance, the selectivity of hidden units in early versus later layers of networks should be studied and compared to recorded neurons in early and later parts of the AC.

Improving Evaluation Measures

A major challenge in sound source separation is simply measuring effectively whether sound source separation has occurred (Vincent et al., 2006). Exactly how one defines good source separation will depend on what purpose the source separation serves. Nevertheless, SDR remains a standard measure used in the field, and there a number of things we could do to use it better. To better evaluate the SDR measure, a baseline SDR metric can be calculated by using the target mixture instead of the target single sound in the SDR definition; then, a true positive sound separation would yield a high SDR score when the predicted separated source is being compared to its target single sound but a low score when compared to a mixture in which it is a component of. Another way to validate that a source separation is genuinely good is to measure the difference between the G1 and G2 predictions in addition to measuring the similarity between the target and predicted source separation. This can be done by incorporating a discriminative term to the SDR that penalizes for similar predictions between G1 and G2. Another important modification could be making separation measures more robust against certain unimportant distortions in the estimation of the source, such as differences between the estimate and the source in overall mean or gain (i.e. shift and scale invariance). Finally, further work should be done to explore why examples that earn high and low source separation scores. Ideally, a low score should correspond to a set of single sources that are hard to separate out due to their similarity, while a high score should correspond to easily separable examples. Such work would not only test the evaluation measure but also help inform model development in helping to elucidate what qualities makes sound separation difficult.

In addition to the source to distortion ratio (SDR), a few other measures are used in supervised source separation research that this model should also use for evaluation: the source to interference ratio (SIR) measures the residual interference from another source in one source's separated output, while the source to artifacts ratio (SAR) captures the artifacts introduced to an output via the sound separation process (Vincent et al., 2006) The triplet of SDR, SIR, and SAR measures are related such that distortion is defined as the sum of interference left and artifacts introduced in source separation. In order to measure how intelligible the separated sounds are, a short time objective intelligibility (STOI) measure, which captures how intelligible denoised speech is, can also be used (Taal et al., 2011).

For datasets like mixtures of speeches on which STOI can not be used, an end-to-end, evaluation pipeline can be developed in which a predicted, source separated output cochleagram is converted back to a sound wave and human listeners are asked to evaluate its intelligibility and/or transcribe the output (Slaney et al., 1994). Then, not only can human evaluation be used but also psychometric data of humans evaluating the output can also be analyzed.

Exploring Alternative Models

In addition to the presented model, which involves regularizing population of hidden units, alternative models should be explored that fit within the unsupervised framework of using subsets of hidden units to perform sound separation. In the case of two sounds forming a mixture, instead of regularizing the hidden layer, two parallel output layers can be distinctly regularized; then, the supervised prediction output would simply be the addition of the two output layers (Figure 22). This output-regularized model is attractive because it corresponds well to the similarity of the sum of the power cochleagrams of single sounds to the power cochleagram of its mixture. For log-compressed cochleagrams, the mixture cochleagram corresponds to the element-wise maximum of its component cochleagrams; thus, the predicted output would simply be the element-wise maximum of the two output layers.

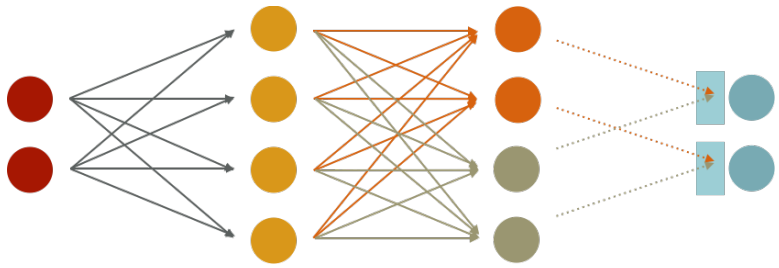


Figure 22: Network Model with Distinguished Output Unit Populations

An alternative model in which the hidden layer is connected to two parallel output unit populations O_1 and O_2 (orange and grey-green circles), which can then be distinctly regularized, and whose sum (for power cochlegrams) or element-wise maximum (for log-compressed cochleagrams) form the predicted output (light blue boxes represent the sum or max function).

Another way to leverage cochleagram compression – either in addition to or instead of distinctly regularizing output units (Figure 22) – is to modify the supervised, predictive task's loss function to explicitly incorporate the way a mixture cochleagram relates to the cochleagram of its individual components. For log-compressed cochleagrams, the loss function can be modified to take the maximum of two output populations as its prediction (Equation 17):

$$\theta^* = \text{argmin}[\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \| y_{nt} - \max(\hat{y}_{O_1(nt)}, \hat{y}_{O_2(nt)}) \|_2^2 + \lambda \| \theta \|_1]$$

Equation 17: Max-Output MLE Predictive Loss with L1 Regularization

where O_1 and O_2 correspond to different output populations whose element-wise maximum forms the output prediction (Figure 22). When training with power cochleagrams, a similar formulation based on the summing effect of mixture cochleagrams can be used (Equation 18):

$$\theta^* = \text{argmin}[\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} \| y_{nt} - (\hat{y}_{O_1(nt)} + \hat{y}_{O_2(nt)}) \|_2^2 + \lambda \| \theta \|_1]$$

Equation 18: Summed-Output MLE Predictive Loss with L1 Regularization

Another possible network is one that predicts not just the expected future but that also parameterizes the distribution over possible future outcomes; for instance, each hidden unit could be defined by two scalar values representing its mean and variance (Bishop, 1994). Reflecting the noise and diversity of variance in real neurons, such a more complete predictive network may produce more robust source separation outputs. The regularizing hidden unit populations model can easily be adapted to this probabilistic network, which would allow further analysis on the difference in variance between selective and non-selective hidden units to be easily conducted.

Finally, while the presented networks are loosely biologically-inspired, they are not biologically-plausible networks because backpropagation – the technique necessary for updating network weights – requires too much precision to feasibly occur in the mammalian brain (Crick, 1989) and also because non-spiking data in the form of cochleagrams is used. Nevertheless, more biologically-plausible networks have been developed (Lillicrap et al., 2014); thus, the application of this dissertation's unsupervised framework and regularization technique to such models should be explored.

Conclusion

In conclusion, while the novel unsupervised source separation framework and hidden unit population regularization technique shows promising results, further research[4] should be done to test its limits and better understand how the regularization technique and resulting networks work. Additionally, alternative network extensions that can utilize the presented unsupervised source separation framework should also be explored.

Nevertheless, this dissertation presents promising proof-of-concept results on the power and biological-consistency of an unsupervised source separation framework that leverages distinctly regularized hidden unit populations.

[4] A few of the aforementioned research directions will be explored, as this project will be continued beyond the MSc project timeline.

Supplementary Figures

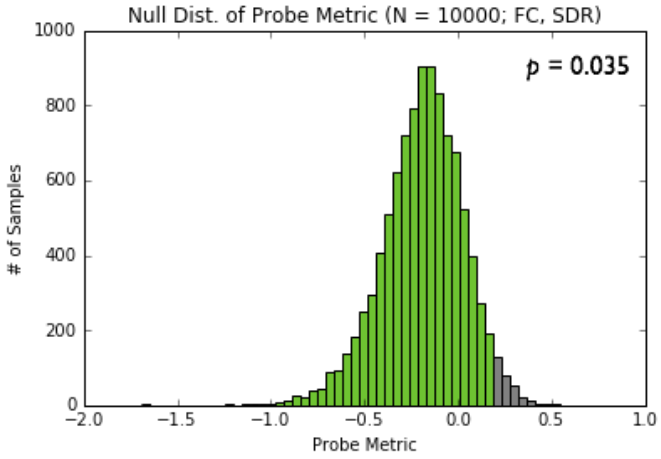


Figure 23: Null Distribution of SDR Probe Metric Generated for Permutation Test (FC)

The FC model's mean SDR score of 0.17 was better than all but 350 of the 10,000 generated samples (p = 0.035). To generate a sample, the 80 hidden units with output connections were randomly split into two groups that were then used to calculate G1 and G2 predictions and the SDR score for each triplet example; the mean SDR score over all triplets for a given hidden unit split was used as the sample. The green bars signify the sampled scores that the model did better than, while the green bars denote sampled scores of random hidden unit splits that outperform the model's explicitly defined G1 and G2 in source separation. The generated distribution has mean, $\mu = -0.19$, and standard deviation, $\sigma = 0.22$.

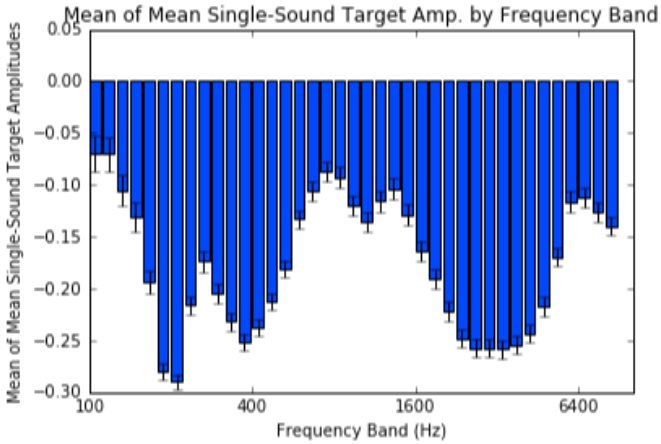


Figure 24: Mean Per-Frequency Band Amplitudes of Single-Sounds in Validation Set

For each non-mixture sound in the validation set, the mean amplitude was calculated for every frequency band, and the mean of the mean amplitudes for all non-mixture sounds was plotted for every frequency. Due to pre-processing data normalization, the mean amplitude across all frequencies is $\mu = -0.18$, with the standard deviation of $\sigma = 0.39$. Notably, activity is not evenly distributed across frequency bands, and there appears to be increased activity around 100 Hz, 800 Hz, and 6400 Hz.

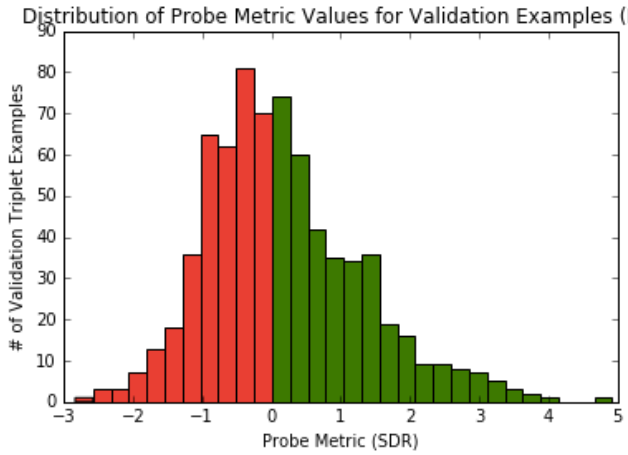


Figure 25: Distribution of SDR Example Probe Scores (FC)

For each triplet example, its SDR score was calculated, and all scores were binned in a histogram ($\mu = 0.17, \theta = 1.13$). 364 of the 720 triplet examples (50.6%) yielded positive SDR scores, which connotes that there is more signal than noise in the optimally predicted source separation; although the distribution has a longer positive tail.

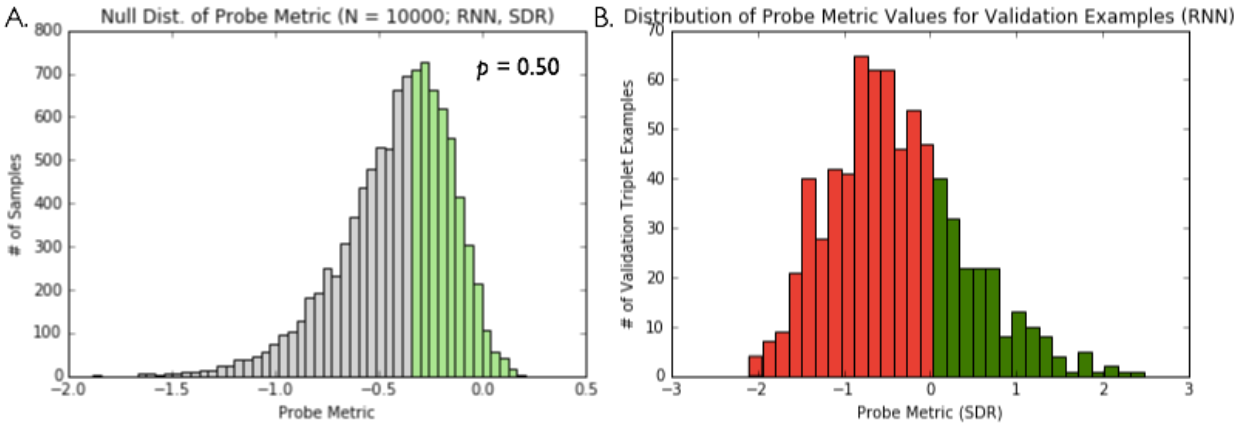


Figure 26: Null Distribution and Model's Distribution of SDR Probe Metric (RNN)

A. *The RNN model's mean SDR score of -0.38 was only better than 5039 of the 1,000 generated samples ($p = 0.50$). Thus, the RNN model's source separation is not significantly better than source separation by randomly assigned hidden group populations. To generate a sample, the 80 hidden units with output connections were randomly split into two groups that were then used to calculate G1 and G2 predictions and the SDR score for each triplet example; the mean SDR score over all triplets for a given hidden unit split was used as the sample. The green bars signify the sampled scores that the model did better than, while the green bars denote sampled scores of random hidden unit splits that outperform the model's explicitly defined G1 and G2 in source separation. The generated distribution has mean, $\mu = -0.42$, and standard deviation, $\sigma = 0.27$. This sub-figure can be compared which contains the same plot generated using the FC model instead.* ***B.*** *For each validation triplet example, its SDR score using the RNN model was calculated, and all scores were binned in a histogram ($\mu = -0.38, \theta = 0.78$). 197 of the 720 triplet examples (27.3%) yielded positive SDR scores, which connotes that there is more signal than noise in the optimally predicted source separation. These two sub-figures can be compared to Figure 23 and Figure 25 respectively, which contain corresponding plots to* ***A*** *and* ***B*** *generated using the FC model instead.*

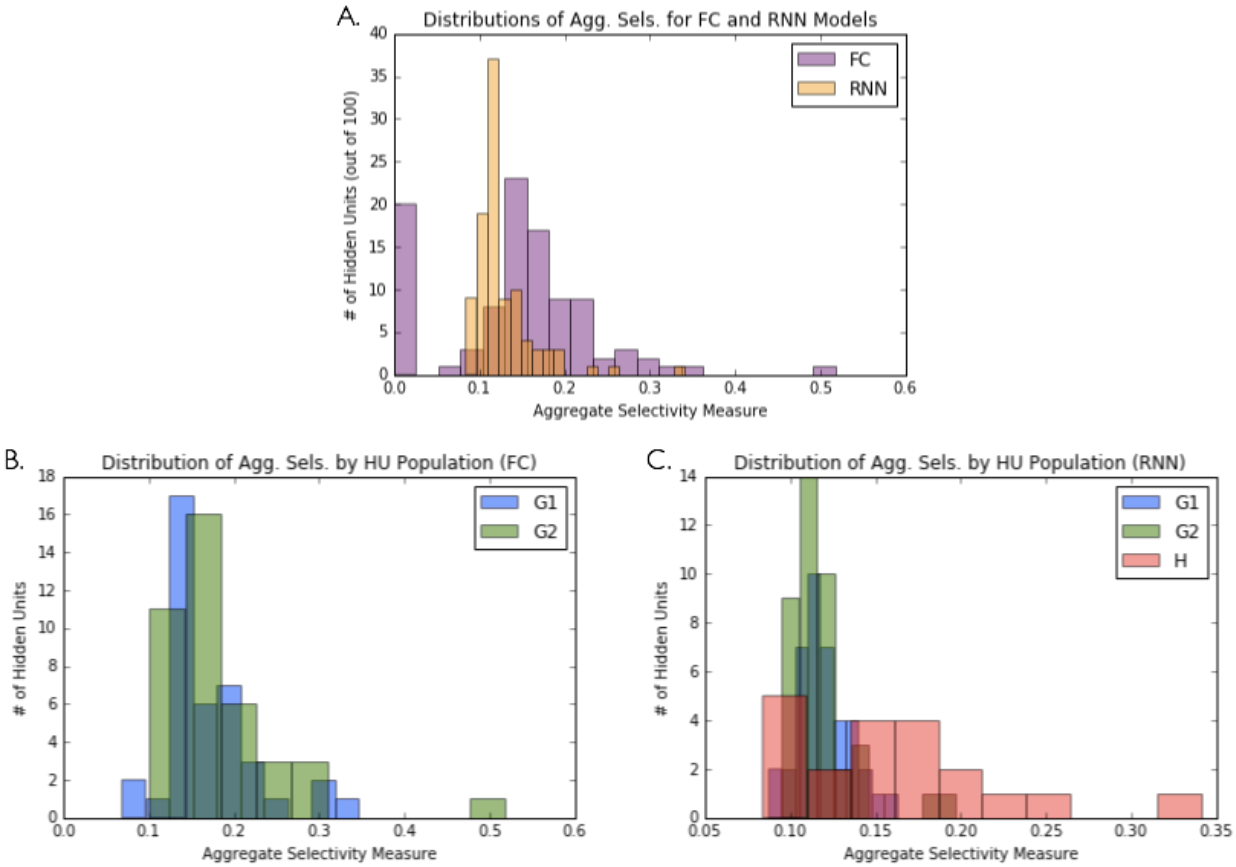


Figure 27: Distribution of Aggregate Selectivity Scores by Model and by Hidden Unit Population

A. *The aggregate selectivity scores of all hidden units for the in-depth-analysis FC ($\mu = 0.14, \theta = 0.09$) and RNN ($\mu = 0.13, \theta = 0.04$) models are binned into histograms. This plot shows that the FC model on average has more selective units than the RNN model, which can also be inferred by examining the top three selective units from each model (Figure 21). 20 FC hidden units have an aggregate selectivity score of 0 because they are dead H units with no output connections.* ***B, C.*** *For the in-depth-analysis FC (**B**) and RNN (**C**) model, the aggregative selectivity scores of relevant hidden unit populations are binned into a histogram. While G1 and G2's selectivity do not appear to vary, the H population has a longer tail of more selective units (**C**). Taking **B** and **C** together, these plots suggest that the RNN's H population influences the selectivity of the whole network via its recurrent connections to G1 and G2 units, while selectivity occurs in the FC model via the increased overall selectivity of all G1 and G2 units.*

References

Al-Rfou R et al. (2016) Theano: A {Python} framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688 Available at: http://arxiv.org/abs/1605.02688.

Bar-Yosef O, Nelken I (2007) The effects of background noise on the neural responses to natural sounds in cat primary auditory cortex. Front Neurosci 1:1–14.

Bee MA, Klump GM, Teki S, Chait M, Kumar S, Kriegstein K Von, Timothy D (2011) Primitive Auditory Stream Segregation : A Neurophysiological Study in the Songbird Forebrain Primitive Auditory Stream Segregation : A Neurophysiological Study in the Songbird Forebrain. J Neurophysiol:1088–1104.

Bell AJ, Sejnowski TJ (1995) An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Comput 7:1129–1159 Available at: http://www.mitpressjournals.org/doi/abs/10.1162/neco.1995.7.6.1129#.VbJaKPmqqko.

Bialek W, Nemenman I, Tishby N (2001) Predictability , Complexity , and Learning. Neural Comput 13:2409–2463.

Bishop CM (1994) Mixture Density Networks. Neural Comput Res Gr Rep 1 Available at: http://www.ncrg.aston.ac.uk.

Bregman AS (1990) Auditory Scene Analysis: The Perceptual Organization of Sound.

Brown GJ, Cooke M (1994) Computational auditory scene analysis. Comput Speech Lang 8:297–336.

Cardoso JF, Laheld BH (1996) Equivariant adaptive source separation. IEEE Trans Signal Process 44:3017–3030.

Comon P (1994) Independent component analysis, A new concept? Signal Processing 36:287–314.

Crick F (1989) The recent excitement about neural networks. Nature 337:129–132 Available at: http://www.nature.com.ezp-prod1.hul.harvard.edu/nature/journal/v337/n6203/pdf/337129a0.pdf\npapers3://publication/uuid/79BF4C49-C77F-4046-968C-9B9A30CE0811.

Dieleman S et al. (2015) Lasagne: First release. Available at: http://dx.doi.org/10.5281/zenodo.27878.

Douglas R, Koch C, Mahowald M, Martin K, Suarez H (1995) Recurrent excitation in neocortical circuits. Science (80-) 269:981–985 Available at: http://science.sciencemag.org/content/269/5226/981.abstract.

Elhilali M, Shamma S a (2008) A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. J Acoust Soc Am 124:3751–3771.

Ellis DPW (1994) A computer implementation of psychoacoustic grouping rules. Proc 12th IAPR Int Conf Pattern Recognition, Vol 2 - Conf B Comput Vis Image Process (Cat No94CH3440-5):108–112 vol.3.

Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. Proc 13th Int Conf Artif Intell Stat 9:249–256 Available at: http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS2010_GlorotB10.pdf.

Harper N, Willmore B, Whittington J, Schnupp J, King A (n.d.) The representation of sound

mixtures in the auditory cortex (in submission).
Heil P, Neubauer H, Irvine DRF (2011) An Improved Model for the Rate-Level Functions of Auditory-Nerve Fibers. J Neurosci 31:15424–15437.
Hill A V. (1910) The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curves. J Physiol 40:iv – vii Available at: http://onlinelibrary.wiley.com/doi/10.1113/jphysiol.1910.sp001386/abstract.
Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Comput 9:1735–1780.
Hsu C-LHC-L, Jang J-SR (2010) On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. IEEE Trans Audio Speech Lang Processing 18:310–319.
Huang P Sen, Kim M, Hasegawa-Johnson M, Smaragdis P (2015) Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation. IEEE/ACM Trans Speech Lang Process 23:2136–2147.
Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9:90–95.
Jang G-J, Lee T-W (2003) A Probabilistic Approach to Single Channel Blind Signal Separation. Adv Neural Inf Process Syst 15:1173–1180 Available at: http://papers.nips.cc/paper/2224-a-probabilistic-approach-to-single-channel-blind-signal-separation.pdf\nfiles/4280/Jang ? Lee - 2003 - A Probabilistic Approach to Single Channel Blind S.pdf\nfiles/4281/2224-a-probabilistic-approach-to-single-channel-blin.
Jones E, Oliphant T, Peterson P, others (n.d.) {SciPy}: Open source scientific tools for {Python}. Available at: http://www.scipy.org/.
Kabal P (2002) TSP speech database. McGill Univ Database Version 0 Available at: http://www-mmsp.ece.mcgill.ca/documents/Downloads/TSPspeech/TSPspeech.pdf.
Kanwal JS, Medvedev A V, Micheyl C (2003) Neurodynamics for auditory stream segregation: tracking sounds in the mustached bat's natural environment. Network-Computation Neural Syst 14:413–435 Available at: <Go to ISI>://WOS:000184986800004.
Kashino K, Tanaka H (1993) A sound source separation system with the ability of automatic tone modeling. Proc Int Comput … Available at: http://www.google.es/search?client=safari&rls=es-es&q=A+sound+source+separation+system+with+the+ability+of+automatic+tone+modeling&ie=UTF-8&oe=UTF-8&redir_esc=&ei=HX3-Tfu9LMLF8QPAyryqCQ\nfile:///Users/julio/Documents/Papers2/1993/Kashino/1993Kashino-A sou.
Kingma D, Ba J (2014) Adam: A Method for Stochastic Optimization. arXiv14126980 [cs]:1–15 Available at: http://arxiv.org/abs/1412.6980\nhttp://www.arxiv.org/pdf/1412.6980.pdf.
Kochkin BS (2002) 10-Year Customer Satisfaction Trends. October.
Le Q V, Jaitly N, Hinton GE (2015) A Simple Way to Initialize Recurrent Networks of Rectified Linear Units. arXiv Prepr arXiv150400941:1–9.
Lillicrap TP, Cownden D, Tweed DB, Akerman CJ (2014) Random feedback weights support learning in deep neural networks. arXiv14110247 [cs, q-bio]:1–27 Available at: http://arxiv.org/abs/1411.0247\nhttp://www.arxiv.org/pdf/1411.0247.pdf.
Lütkenhöner B (2008) Threshold and beyond: Modeling the intensity dependence of

auditory responses. JARO - J Assoc Res Otolaryngol 9:102–121.
Mesgarani N, Chang EF (2013) Selective cortical representation of attended speaker in multi- talker speech perception. 10:54–56.
Micheyl C, Tian B, Carlyon RP, Rauschecker JP (2005) Perceptual organization of tone sequences in the auditory cortex of awake macaques. Neuron 48:139–148.
Mitchell T, Cohen W, Hruschka E, Talukdar P, Betteridge J, Carlson A, Dalvi B, Gardner M (2015) Never-Ending Learning. Proc Conf Artif Intell.
Morrison GS, Zhang C, Enzinger E, Ochoa F, Bleach D, Johnson M, Folkes BK, De Souza S, Cummins N, Chow D (2015) Forensic database of voice recordings of 500+ Australian English speakers. Available at: http://databases.forensic-voice-comparison.net/.
O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cereb Cortex 25:1697–1706.
Roweis SST (2001) One microphone source separation. Adv Neural Inf Process Syst:793–799 Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.3986&rep=rep1&type=pdf.
Salisbury J, Palmer SE (2015) Optimal prediction and natural scene statistics in the retina. :1–11 Available at: http://arxiv.org/abs/1507.00125.
Singer Y, Teramoto Y, Willmore B, King A, Schnupp J, Harper N (n.d.) The neural code of sensory cortex is optimised to predict future input (in submission).
Slaney M, Naar D, Lyon RF (1994) Auditory Model Inversion for Sound Separation. Proc ICASSP Available at: http://www.ncbi.nlm.nih.gov/pubmed/15003161\nhttp://cid.oxfordjournals.org/lookup/doi/10.1093/cid/cir991\nhttp://www.scielo.cl/pdf/udecada/v15n26/art06.pdf\nhttp://www.scopus.com/inward/record.url?eid=2-s2.0-84861150233&partnerID=tZOtx3y1.
Taal CH, Hendriks RC, Heusdens R, Jensen J (2011) An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans Audio, Speech Lang Process 19:2125–2136.
van der Walt S, Colbert SC, Varoquaux G (2011) The NumPy Array: A Structure for Efficient Numerical Computation. Comput Sci Eng 13:22–30.
Victor Z, Seneff S, Glass J (1990) TIMIT acoustic-phonetic continuous speech corpus. Speech Commun 9:351–356 Available at: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1.
Vincent E, Gribonval R, Févotte C (2006) Performance measurement in blind audio source separation. IEEE Trans Audio, Speech Lang Process:1462–1469.
Wang D, Brown GJ (2006) Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Available at: http://www.lavoisier.fr/livre/notice.asp?id=OKLWRLAKRO6OWH.
Willmore BDB, Schoppe O, King AJ, Schnupp JWH, Harper NS (2016) Incorporating Midbrain Adaptation to Mean Sound Level Improves Models of Auditory Cortical Processing. J Neurosci 36:280–289 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26758822\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4710761.
Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR,

Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” Neuron 77:980–991 Available at: http://dx.doi.org/10.1016/j.neuron.2012.12.037.