
Neural Separation of Observed and Unobserved Distributions

Tavi Halperin¹ Ariel Ephrat² Yedid Hoshen^{1,3}

Abstract

Separating mixed distributions is a long standing challenge for machine learning and signal processing. Most current methods either rely on making strong assumptions on the source distributions or rely on having training samples of each source in the mixture. In this work, we introduce a new method—Neural Egg Separation—to tackle the scenario of extracting a signal from an unobserved distribution additively mixed with a signal from an observed distribution. Our method iteratively learns to separate the known distribution from progressively finer estimates of the unknown distribution. In some settings, Neural Egg Separation is initialization sensitive, we therefore introduce Latent Mixture Masking which ensures a good initialization. Extensive experiments on audio and image separation tasks show that our method outperforms current methods that use the same level of supervision, and often achieves similar performance to full supervision.

1. Introduction

Humans are remarkably good at separating data coming from a mixture of distributions, e.g. hearing a person speaking in a crowded cocktail party. Artificial intelligence, on the other hand, is far less adept at separating mixed signals. This is an important ability as signals in nature are typically mixed, e.g. speakers are often mixed with other speakers or environmental sounds, objects in images are typically seen along other objects as well as the background. Understanding mixed signals is harder than understanding pure sources, making source separation an important research topic.

Most previous work focused on the following settings:

Full supervision: The learner has access to a training set

¹Department of Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel ²Google Research ³Facebook AI Research. Correspondence to: Tavi Halperin <tavi.halperin@mail.huji.ac.il >.

including samples of mixed signals $\{y\} \in \mathcal{Y}$ as well as the ground truth sources of the same signals $\{b\} \in \mathcal{B}$ and $\{x\} \in \mathcal{X}$ (such that $y = x + b$). Having such strong supervision is very potent, allowing the learner to directly learn a mapping from the mixed signal y to its sources (x, b) . Obtaining such strong supervision is nearly never possible, as it requires knowing for each input mixture y , its exact separate signals (x, b) . This information is rarely available in a single microphone setting.

Synthetic full supervision: The learner has access to a training set containing samples from the mixed signal $\{y\} \in \mathcal{Y}$ as well as samples from all source distributions $\{b\} \in \mathcal{B}$ and $\{x\} \in \mathcal{X}$. The learner however does not have access to paired sets of the mixed and unmixed signal ground truth (that is for any given y in the training set, its separate b and x are unknown). This supervision setting is more realistic than the fully supervised case, and occurs when each of the source distributions can be sampled in isolation (e.g. we can record a violin and piano separately in a studio and can thus obtain unmixed samples of each of their distributions). It is typically solved by learning to separate synthetic mixtures $b + x$ of independently sampled b and x .

No supervision: The learner only has access to training samples of the mixed signal \mathcal{Y} but not to sources \mathcal{B} and \mathcal{X} . Although this settings puts the least requirements on the training dataset, it is a hard problem and can be poorly specified in the absence of strong assumptions and priors. It is generally necessary to make strong assumptions on the properties of the component signals (e.g. smoothness, low rank, periodicity) in order to make progress in separation. This limits the applicability of such methods.

In this work we concentrate on the *semi-supervised* setting: unmixing of signals in the case where the mixture \mathcal{Y} consists of a signal coming from an unobserved distribution \mathcal{X} and another signal from an observed distribution \mathcal{B} (i.e. the learner has access to a training set of clean samples such that $\{b\} \in \mathcal{B}$ along with different mixed samples $\{y'\} \in \mathcal{Y}$). One possible way of obtaining such supervision is to label every element of a signal (e.g. short waveform segment) by a label, indicating if it comes only from the observed distribution \mathcal{B} or if it is a mixture of both distributions $\mathcal{B} + \mathcal{X}$. The task is to learn a parametric function able to separate the mixed signal $y \in \mathcal{Y}$ into sources $x \in \mathcal{X}$ and $b \in \mathcal{B}$ s.t.

$y = b + x$. Such supervision is more generally available than full supervision, while the separation problem becomes much simpler than when fully unsupervised.

We introduce a novel method: Neural Egg Separation (NES), consisting of iterative: *i*) estimation of samples from the unobserved distribution \mathcal{X} *ii*) synthesis of mixed signals from known samples of \mathcal{B} and estimated samples of \mathcal{X} *iii*) training of separation function for the mixed signal. Iterative refinement of the estimated samples of \mathcal{X} significantly increases the accuracy of the learned separation function. The method is named Neural Egg Separation, as it is akin to the iterative technique commonly used for separating egg whites and yolks.

As an iterative technique, NES can be initialization sensitive. We therefore introduce another method — Latent Mixture Masking (LMM) — to provide NES with a strong initialization. Our method trains two deep generators end-to-end using Latent Mixtures to model the observed and unobserved sources (\mathcal{B} and \mathcal{X}). we found that a simple initialization is sufficient when \mathcal{X} and \mathcal{B} are uncorrelated, whereas LMM-initialization is most important when \mathcal{X} and \mathcal{B} are strongly correlated such as e.g. separation of music into instruments and vocals. Initialization by LMM was found to be much more effective than by adversarial methods.

Experiments are conducted across multiple domains (image, music, voice) validating the effectiveness of our method, and its superiority over current methods that use the same level of supervision. Our semi-supervised method is often competitive with the fully supervised baseline, while making few assumptions on the nature of the component signals and requiring lightweight supervision. An analysis of the assumptions made by the method is detailed in Sec. 5.

2. Previous Work

Source separation: Separation of mixed signals has been extensively researched. In this work, we focus on single channel separation. Unsupervised (blind) single-channel methods include: Robust Principal Component Analysis (RPCA) (Huang et al., 2012) and Single-channel Independent Component Analysis (ICA) (Davies & James, 2007). These methods attempt to use coarse priors about the signals such as low rank, sparsity or non-gaussianity. Hidden Markov Models (HMM) can be used as a temporal prior for longer clips (Roweis, 2001), however here we do not assume long clips. Supervised source separation has also been extensively researched, classic techniques often used learned dictionaries for each source e.g. Non-negative Matrix Factorization (NMF) (Wilson et al., 2008). Recently, neural network-based separation gained popularity, usually learning a regression between the mixed and unmixed signals either directly (Huang et al., 2014), or by regressing

a multiplicative mask (Wang et al., 2014; Yu et al., 2017). Some methods were devised to exploit the temporal nature of long audio signal by using Recurrent Neural Networks (RNNs) (Mimilakis et al., 2017), in this work we concentrate on separation of short audio clips and consider such line of works as orthogonal. One related direction is Generative Adversarial Source Separation (Stoller et al., 2017; Subakan & Smaragdis, 2017) that uses adversarial training to match the unmixed source distributions. This is needed to deal with correlated sources for which learning a regressor on synthetic mixtures is less effective. We present an Adversarial Masking (AM) method that tackles the semi-supervised rather than the fully supervised scenario and overcomes mixture collapse issues not present in the fully supervised case. We found that non-adversarial methods perform better for the initialization task.

The most related set of works is semi-supervised audio source separation (Smaragdis et al., 2007; Barker & Virtanen, 2014), which like our work attempt to separate mixtures \mathcal{Y} given only samples from the distribution of one source \mathcal{B} . Typically NMF or Probabilistic Latent Component Analysis (PLCA) (which is a similar algorithm with a probabilistic formulation) are used. We show experimentally that our method significantly outperforms NMF. A very early related technique is Spectral Subtraction (Boll, 1979), however it can only handle very simple unknown sources.

Disentanglement: Similarly to source separation, disentanglement also deals with separation in terms of creating a disentangled representation of a source signal, however its aim is to uncover latent factors of variation in the signal such as style and content or shape and color e.g. (Denton et al., 2017; Higgins et al., 2016). Differently from disentanglement, our task is separating signals rather than the latent representation.

Generative Models: Generative models learn the distribution of a signal directly. Classical approaches include: singular-value decomposition (SVD) for general signals and NMF (Lee & Seung, 2001) for non-negative signals. Recently several deep learning approaches dominated generative modeling including: Generative Adversarial Network (GAN) (Goodfellow et al., 2016), Variational Autoencoder (VAE) (Kingma & Welling, 2013) and Generative Latent Optimization (GLO) (Bojanowski et al., 2018). Adversarial training (for GANs) is rather tricky and often leads to mode-collapse. GLO is non-adversarial and allows for direct latent optimization for each source making it more suitable than VAE and GAN.

3. Neural Egg Separation (NES)

In this section we present our method for separating a mixture of sources of known and unknown distributions. We

denote the mixture samples y , the corresponding samples with the observed distribution b and the samples from the unobserved distribution x . Our objective is to learn a parametric function $T(y)$, such that $b = T(y)$.

Full Supervision: In the fully supervised setting, where matching pairs (y, b) are available, this task reduces to a standard supervised regression problem, in which a parametric function $T(y)$ (typically a deep neural network) is used to directly optimize:

$$T = \arg \min_{T'} \sum_{(y,b)} L_1(T'(y), b) \quad (1)$$

Mixed-unmixed pairs are usually unavailable, but in some cases it is possible to obtain a training set which includes independent samples from \mathcal{X} and \mathcal{B} e.g. (Wang et al., 2014; Yu et al., 2017). Methods typically randomly sample $x \in \mathcal{X}$ and $b \in \mathcal{B}$ sources and synthetically create mixtures $y' = x + b$. The synthetic pairs (y', b) can then be used to optimize Eq. 1. Note that in cases where \mathcal{X} and \mathcal{B} are correlated (e.g. vocals and instrumental accompaniment which are temporally dependent), random synthetic mixtures of x and b might not be representative of \mathcal{Y} and cause difficulty generalizing to real mixtures.

Semi-Supervision: In many scenarios, clean samples of both mixture components are not available. Consider for example a street musical performance. Although crowd noises without street performers can be easily observed, street music without crowd noises are much harder to come by. In this case, therefore, samples from the distribution of crowd noise \mathcal{B} are available, whereas the samples from the distribution of the music \mathcal{X} are unobserved. Samples from the distribution of the mixed signal \mathcal{Y} i.e. the crowd noise mixed with the musical performance are also available.

The example above illustrates a class of problems for which the distribution of the mixture and a single source are available, but the distribution of another source is unknown. In such cases, it is not possible to optimize Eq. 1 directly due to the unavailability of paired (y, b) .

Neural Egg Separation: Fully-supervised optimization (as in Eq. 1) is very effective when pairs of (y, b) are available. We present a novel algorithm, which iteratively solves the semi-supervised task as a sequence of supervised problems without any clean training examples of \mathcal{X} .

The core idea of our method is that although no clean samples from \mathcal{X} are given, it is still possible to learn to separate mixtures of observed samples b from distribution \mathcal{B} combined with some estimates of the unobserved distribution samples x^t (where t denotes the iteration of NES). Synthetic mixtures are created by randomly sampling an approximate sample x^t from the unobserved distribution and combining with training sample b , thereby creating pairs (y^t, b) for

Algorithm 1 Neural Egg Separation (NES)

Input: samples of: mixture $\{y\}$, observed source $\{b\}$
Initialize: synthetic unobservable samples with $x^0 \leftarrow c \cdot y$ or using AM or LMM
while $t < N$ **do**
 Initialize $T()$ with random weights
 Synthesize mixtures $y^t = b + x^t$ for all b in \mathcal{B} with randomly sampled x^t
 Optimize separation function for P epochs: $T^{t+1} = \arg \min_{T'} \sum_{(y^t,b)} L_1(T'(y^t), b)$
 Update estimates of unobserved distribution samples: $x^{t+1} = y - T^{t+1}(y)$
end while

supervised training:

$$y^t = b + x^t \quad (2)$$

Note that the empirical distribution of synthetic mixtures \mathcal{Y}^t might differ from the real mixture sample distribution \mathcal{Y} . We show empirically that there are interesting cases for which it converges towards the correct distribution: $\mathcal{Y}^t \rightarrow \mathcal{Y}$.

During each iteration of NES, a neural separation function $T^{t+1}(y^t)$ is trained on the created pairs by optimizing the following term:

$$T^{t+1} = \arg \min_{T'} \sum_{(y^t,b)} L_1(T'(y^t), b) \quad (3)$$

At the end of each iteration, the separation function $T^t()$ can be used to approximately separate the training mixture samples y into their sources:

$$x^t = y - T^t(y) \quad (4)$$

The refined samples $x^t \in \mathcal{X}^t$ are used for creating synthetic pairs for training $T^{t+1}(y^t)$ in the next iteration (as in Eq. 3).

The above method relies on having an estimate of the unobserved distribution samples as input to the first iteration (\mathcal{X}^0). One simple scheme is to initialize the estimates of the unobserved distribution samples in the first iteration as $x^0 = c \cdot y$, where c is a constant fraction (typically 0.5). Although this initialization is very naive, we show that it performs well where the sources are independent. More advanced initializations will be discussed below.

At test time, separation is carried out by applying the trained separation function $T()$ (exactly as in Eq. 4).

Our full algorithm is described in Alg. 1. For optimization, we use SGD using ADAM update with a learning rate of

0.001. In total we perform $N = 10$ iterations, each consisting of optimization of T and estimation of x^t , $P = 25$ epochs are used for each optimization of Eq. 3.

Latent Mixtures: NES is an iterative method and relies on having a good initialization. It does not take into account correlation between \mathcal{X} and \mathcal{B} e.g. vocals and instrumental tracks are highly related, whereas randomly sampling pairs of vocals and instrumental tracks is likely to synthesize mixtures quite different from \mathcal{Y} .

We present our method Latent Mixtures (LM), which separates mixtures by a distributional constraint enforced via latent generative modeling of the source signals. The method uses some latent optimization ideas from GLO (Bojanowski et al., 2018). The novelty of LM is using mixtures of GLO models for separation, which has not been done before. LM training consists of two stages. We first learn a generator $G_B()$, with which for every observed training sample b from \mathcal{B} , a latent code z_b can be found such that b is reconstructed by the generator: $b = G_B(z_b)$. We learn end-to-end both the parameters of the generator $G_B()$ as well as a latent code z_b for every training sample b . The per-sample latent codes are found by direct gradient descent over the values of z_b (similar to word embeddings), rather than by a feedforward encoder. This stage is equivalent to GLO. The optimization is given by:

$$\arg \min_{z_b, G_B} \sum_{b \in \mathcal{B}} \ell(G(z_b), b) \quad (5)$$

Given the learned generator $G_B(z)$ for the \mathcal{B} distribution, we learn generator $G_X(z)$ for the unobserved distribution \mathcal{X} . The idea is that every \mathcal{Y} domain training sample y , is described by mixing a \mathcal{B} domain signal generated by $G_B(z_y^B)$ as well as a \mathcal{X} domain signal generated by $G_X(z_y^X)$. Note that we do not know the actual source b but only have a generative model prior for the \mathcal{B} distribution. As we have already learned $G_B()$ in the previous stage, we just need to learn $G_X()$ as well as the per y sample latent codes z_y^X and z_y^B . The optimization function is therefore:

$$\arg \min_{z_y^X, z_y^B, G_X} \ell(G_B(z_y^B) + G_X(z_y^X), y) \quad (6)$$

Similarly to (Bojanowski et al., 2018), we found that forcing latent codes to lay in a unit ball provides important regularization. We use $\ell = L_1$ except for color images, where we found it advantageous to use a VGG perceptual loss (implementation taken from (Hoshen & Wolf, 2018)).

Once $G_B()$ and $G_X()$ are trained, we infer the latent codes for a test mixture by:

$$\arg \min_{z_y^X, z_y^B} \ell(G_B(z_y^B) + G_X(z_y^X), y) \quad (7)$$

Our estimate for the sources is then:

$$\tilde{b} = G_B(z_y^B) \quad \tilde{x} = G_X(z_y^X) \quad (8)$$

Masking Function: In additive separation tasks, the mixed signal y is the sum of two positive signals x and b . Instead of synthesizing the new sample, we can learn a neural network separation mask $m(y)$ that specifies the fraction of the signal which comes from \mathcal{B} at each pixel. The attractive feature of the mask is always being in the range $[0, 1]$ (in the case of positive additive mixtures of signals). Even a constant mask will preserve all signal gradients (at the cost of introducing spurious gradients). Mathematically this can be written as:

$$T(y) = y \odot m(y) \quad (9)$$

For NES (and baseline AM described below), we implement the mapping function $T(y)$ using the element-wise product of the masking function and the mixture signal: $y \odot m(y)$. In practice, we find that learning a masking function yields much better results than synthesizing the signal directly (in line with other works e.g. (Wang et al., 2014; Gabbay et al., 2017)).

LM does not provide a way for learning the mask directly. We refine its estimate by computing an effective mask from the element-wise ratio of estimated sources. We name the combination of LM and the post-processing masking operation, Latent Mixture Masking (LMM):

$$m_{LMM}(y) = \frac{G_B(z_y^B)}{G_B(z_y^B) + G_X(z_y^X)} \quad (10)$$

Initializing Neural Egg Separation by LMM: We devise the following method: *i*) Train LMM on the training set and infer the mask for each mixture. This is operated on images or mel-scale spectrograms at 64×64 resolution *ii*) For audio: upsample the mask to the resolution of the high-resolution linear spectrogram and compute an estimate of the \mathcal{X} source linear spectrogram on the training set *iii*) Run NES on the observed \mathcal{B} and estimated \mathcal{X} . We find experimentally that this initialization scheme improves NES to the point of being competitive with fully-supervised training.

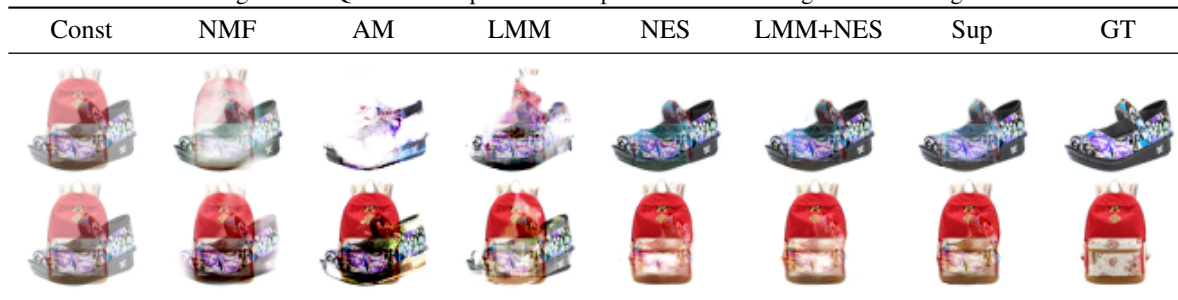
4. Experiments

To evaluate the performance of our method, we conducted experiments on distributions taken from multiple real-world domains: images, speech and music, in cases where the two signals are correlated and uncorrelated.

We evaluated our method against 3 baseline methods:

Constant Mask (Const): This baseline uses the original mixture as the estimate.

Figure 1. A Qualitative Separation Comparison on Mixed Bag and Shoe Images



Semi-supervised Non-negative Matrix Factorization (SS-NMF): This baseline method, proposed by (Smaragdīs et al., 2007), first trains a set of l bases on the observed distribution samples B by Sparse NMF (Hoyer, 2004; Kim & Park, 2007). It factorizes $B = H_b W_b$, with activations H_b and bases W_b , all matrices are non-negative. The optimization is solved using the Non-negative Least Squares solver by (Kim & Park, 2011). It then proceeds to train another factorization on the mixture Y training samples with $2l$ bases, where the first l bases (W_b) are fixed to those computed in the previous stage: $Y = H_y^b * W_b + H_y^x * W_x$. The separated sources are then: $\tilde{x} = H_y^x * W_x$ and $\tilde{b} = H_y^b * W_b$.

Adversarial Masking (AM): As an additional contribution, we introduce a new adversarial semi-supervised method, to improve over the shallow NMF baseline. AM trains a masking function $m(\cdot)$ so that after masking, the training mixtures are indistinguishable from the distribution of source \mathcal{B} under an adversarial discriminator $D(\cdot)$. The loss functions (using LS-GAN (Mao et al., 2017)) are given by:

$$D = \arg \min_{D'} \sum_{y \in \mathcal{Y}} D'(y \odot m(y))^2 + \sum_{b \in \mathcal{B}} (D'(b) - 1)^2 \quad (11)$$

$$m = \arg \min_{m'} \sum_{y \in \mathcal{Y}} (D(y \odot m'(y)) - 1)^2 \quad (12)$$

Differently from CycleGAN (Zhu et al., 2017) and DiscoGAN (Kim et al., 2017), AM is not bidirectional and cannot use cycle constraints. We have found that adding a magnitude prior $L_1(m(y), 1)$ improves performance and helps prevent collapse. To partially alleviate mode collapse, we use Spectral Norm (Miyato et al., 2018) on the discriminator.

We evaluated our proposed methods:

Latent Mixture Masking (LMM): LMM on mel-spectrograms or images at 64×64 resolution.

Neural Egg Separation (NES): The NES method detailed in Sec. 3. Initializing \mathcal{X} estimates using a constant (0.5) mask over \mathcal{Y} training samples.

Initialization by Another Method (AM+NES and LMM+NES): Initializing NES with the \mathcal{X} estimates obtained by Adversarial Masking or by Latent Mixture Masking.

To upper bound the performance of our method, we also compute a *fully supervised* baseline, for which paired data ($y = x + b, b$) of $b \in \mathcal{B}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ are available. We train a masking function with the same architecture as used by all other regression methods to directly regress synthetic mixtures to unmixed sources. This method uses more supervision than our method and is an upper bound.

Please see the appendix for elaborate implementation details.

4.1. Separating Mixed Images

MNIST We evaluate our method on image separation using the following experimental protocol. We split the MNIST dataset (LeCun & Cortes, 2010) into two classes, the first consisting of the digits 0-4 and the second consisting of the digits 5-9. We conduct experiments where one source has an observed distribution \mathcal{B} while the other source has an unobserved distribution \mathcal{X} . We use $12k$ \mathcal{B} training images as the \mathcal{B} training set, while for each of the other $12k$ \mathcal{B} training images, we randomly sample a \mathcal{X} image and additively combine the images to create the \mathcal{Y} training set. We evaluate the performance of our method on 5000 \mathcal{Y} images similarly created from the test set of \mathcal{X} and \mathcal{B} . The experiment was repeated for both directions i.e. 0-4 being \mathcal{B} while 5-9 in \mathcal{X} , as well as 0-4 being \mathcal{X} while 5-9 in \mathcal{B} .

In Tab. 1, we report our results on this task. For each experiment, the top row presents the results (peak signal-to-noise Ratio (PSNR) and structural similarity (SSIM)) on the \mathcal{X} test set. Due to the simplicity of the dataset, NMF achieved reasonable performance on this dataset. LMM achieves better SSIM but worse PSNR than NMF while AM performed 1-2dB better. NES achieves much stronger performance than all other methods, achieving about 1dB worse than the fully supervised performance. Initializing NES with the masks obtained by LMM, results in similar performance

to the fully-supervised upper bound. Initialization by AM achieved similar but slightly inferior performance to initialization by LMM and were omitted from the table for clarity.

Bags and Shoes To evaluate our method on more realistic images, we evaluate on separating mixtures consisting of pairs of images sampled from the Handbags (Zhu et al., 2016) and Shoes (Yu & Grauman, 2014) datasets, which are commonly used for evaluation of conditional image generation methods. To create each \mathcal{Y} mixture image, we randomly sample a shoe image from the Shoes dataset and a handbag image from the Handbags dataset and sum them. For the observed distribution, we sample another 5000 different images from a single dataset. We evaluate our method both for cases when the \mathcal{X} class is Shoes and when it is Handbags.

From the results in Tab. 1, we can observe that NMF failed to preserve fine details, penalizing its performance metrics. LMM (which used a VGG perceptual loss) performed much better, due to greater expressiveness. AM performance was similar to LMM on this task, as the perceptual loss and stability of training of non-adversarial models helped LMM greatly. NES performed much better than all other methods, even when initialized from a constant mask. Initialization by LMM, helped NES achieve stronger performance, nearly identical to the fully-supervised upper bound. It performed better than initialization by AM (not shown in table) which achieved 22.5/0.85 and 22.7/0.86. Similar conclusions can be drawn from the qualitative comparison in the figure above.

We tested our method on a standard denoising task, where the observed source is clean images and noise unobserved. We use positively clamped Gaussian noise $\sigma = 0.1$. For methods GLO/ours/supervised, we obtained PSNR: 24.4/28.5/28.5 SSIM: 0.76/0.88/0.88 on the Bags dataset and PSNR: 25.2/29.4/29.4 SSIM: 0.83/0.9/0.9 on the Shoes dataset. Our method seems to be well suited for denoising.

4.2. Separating Speech and Environmental Noise

Separating environmental noise from speech is a long standing problem in signal processing. Although supervision for both human speech and natural noises can generally be obtained, we use this task as a benchmark to evaluate our method’s performance on audio signals where \mathcal{X} and \mathcal{B} are not dependent. This benchmark is a proxy for tasks for which a clean training set of sounds from \mathcal{X} cannot be obtained e.g. for animal sounds in the wild, where background sounds without animal noises can easily be recorded, but clean sounds made by the animal with no background sounds are unlikely to be available.

For our experiments, we use the Oxford-BBC Lip Reading in the Wild (LRW) Dataset (Chung & Zisserman, 2016)

for speech. For noise we use audio segments from ESC-50 (Piczak, 2015), a dataset of environmental audio recordings organized into 50 semantic classes. Detailed description of our implementation for audio pre-processing and mask training can be found in the appendix. Separation quality is measured using Signal-to-Distortion ratio (SDR), measured using the BSS Eval toolbox (Vincent et al., 2006; Stöter et al., 2018).

From the speech results in Tab. 2, we can observe that LMM performed similarly to Semi-Supervised NMF, and AM training performed about 3dB better than LMM. Due to the independence between the sources in this task, NES performed well, even when trained from a constant mask initialization. Performance was very close to the fully supervised result (when speech is unobserved). In this setting, initializing NES with the speech estimates obtained by LMM (or AM) did not yield improved performance.

We present in Fig. 2 the results of the different methods on a mixture from the speech dataset. It can be observed that LM captures the general features of the sources, but is not able to exactly capture fine detail. The masking operation in LMM helps it recover more fine-grained details, and results in much cleaner separations. We observe that NES converged quite quickly, and results improve further with increasing iterations. Quantitative SDR results are in line with this finding. A graph of SDR for different $NES(k)$ is presented in the appendix.

4.3. Music Separation

Separating music into singing voice and instrumental music as well as drums separation from instrumental music has been a standard task for the signal processing community. Here our objective is to understand the behavior of our method in settings where \mathcal{X} and \mathcal{B} are dependent.

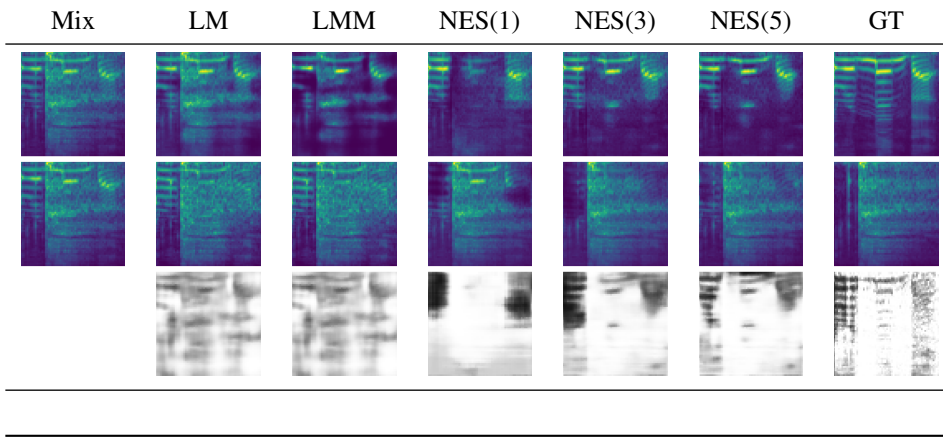
We use the MUSDB18 Dataset (Rafii et al., 2017), consisting for each music track of separate signal streams for the mixture, drums, bass, the rest of the accompaniment, and the vocals. We convert the audio tracks to mono, resample to 20480 Hz, and then follow the same procedure as for speech to obtain input audio features.

From the music results in Tab. 2, we can observe that NMF was the worst performer in this setting (as its simple bases do not generalize well between songs). LMM was able to do much better than NMF and was even competitive with NES on vocal-instrumental separation. Due to the dependence between the two sources and low SNR, initialization proved important for NES. Constant initialization NES performed similarly to AM and LMM. Initialization NES by LMM masks performed much better than all other methods and was competitive with the supervised baseline. LMM initialization was better than AM initialization.

Table 1. Image Separation Accuracy (PSNR dB/SSIM)

\mathcal{X}	\mathcal{B}	Const	NMF	AM	LMM	NES	LMM+NES	Supervised
0-4	5-9	10.6/0.65	16.5/0.71	17.8/0.83	15.1/0.76	23.4/ 0.95	23.9/0.95	24.1/0.96
5-9	0-4	10.8/0.65	15.5/0.66	18.2/0.84	15.3/0.79	23.4/ 0.95	23.8/0.95	24.4/0.96
Bags	Shoes	6.9/0.48	13.9/0.48	15.5/0.67	15.1/0.66	22.3/0.85	22.7/0.86	22.9/0.86
Shoes	Bags	10.8/0.65	11.8/0.51	16.2/0.65	14.8/0.65	22.4/0.85	22.8/0.86	22.8/0.86

Figure 2. A qualitative comparison of mixtures of speech and noise (top and middle rows, respectively) separated by LM and LMM, as well as NES after k iterations. NES(k) denotes NES after k iterations. Note that LM and LMM share the same mask (bottom row), since LMM is generated by the mask computed from LM.



5. Understanding the Limits of NES

This section will investigate a few scenarios under which NES is expected to converge.

Optimal Masking: At each iteration, we solve the following optimization problem: $\mathcal{L}_{NES}^{t+1} = L_1(m^{t+1}(y^t) \odot y^t, b)$, where t is the iteration number, y^t is a synthetic mixture consisting of the sum of a random observed sample $b \in \mathcal{B}$ and an estimated sample $x^t \in \mathcal{X}^t$ ($y^t = b + x^t$). At iteration t , the optimal mask $m^{t+1}(\cdot)$ is:

$$m^{t+1}(y^t) = \frac{b}{y^t} \quad (13)$$

There are several requirements for this optimization to work: *i*) Similarly to other learning-based source separation models, we assume that every mixture y has a unique decomposition into separate sources $x \in \mathcal{X}$ and $b \in \mathcal{B}$. This means that the sources need to have distinct forms. This assumption is not exactly satisfied in practice, but there are cases where it is a good approximation. *ii*) We assume that the optimization method is able to find the optimal solution—despite the non-convexity of the network. The network needs to be sufficiently large to fit the data. These requirements are shared by most other deep learning works.

Generalization from \mathcal{Y}^t to \mathcal{Y} : The objective of source separation (in our formulation) is to learn the mask yield-

ing $m(y) = \frac{b}{y}$ for every $y \in \mathcal{Y}$. NES, however, is only able to operate on the approximate distribution \mathcal{Y}^t , where t is the iteration number. To achieve the supervised performance, NES attempts to progressively improve its estimation: $\mathcal{Y}^t \rightarrow \mathcal{Y}$. At each iteration, the approximation of y is updated using the most recent mask: $y^{t+1} = b + (y - m^{t+1}(y)y)$ (This approximation is not actually known to us as the particular b component of y is unknown). Convergence of the distribution can be measured by the difference between the estimated y^t and the correct sample y . The absolute error $|e^{t+1}(y)|$ is defined below (in this discussion all operations are element-wise):

$$|e^{t+1}| = |y^{t+1} - y| = |b - m^{t+1}(y) \odot y| \quad (14)$$

As m^{t+1} was trained on \mathcal{Y}^t , rather than \mathcal{Y} , we do not know a priori how it generalizes on the true \mathcal{Y} distribution. Let us consider several scenarios:

Perfect Generalization: If m^{t+1} trained on \mathcal{Y}^t generalizes perfectly to all $y \in \mathcal{Y}$, then $m^{t+1}(y) = \frac{b}{y}$. In this case, $|e^{t+1}(y)| = |b - \frac{b}{y} \odot y| = 0$. A single iteration is therefore sufficient for convergence.

Locally Invariant m^{t+1} : Instead of assuming perfect generalization, we consider the case that m^{t+1} is locally invariant around y^t values. The assumption is that $m^{t+1}(y) \simeq$

Table 2. Audio Separation Accuracy (Median SDR dB)

\mathcal{X}	\mathcal{B}	<i>Const</i>	<i>NMF</i>	<i>AM</i>	<i>LMM</i>	<i>NES</i>	<i>AM+NES</i>	<i>LMM+NES</i>	<i>Supervised</i>
Vocals	Instrumental	0.0	0.0	0.0	0.6	0.3	1.2	2.0	2.8
Drums	Instrumental	0.1	-0.6	1.2	0.8	1.3	2.9	3.4	3.7
Speech	Noise	3.0	2.7	6.0	2.3	7.8	7.2	7.6	8.0
Noise	Speech	3.0	2.8	5.2	2.7	6.3	6.4	6.1	8.1

$m^{t+1}(y^t) = \frac{b}{y^t}$. The error will become:

$$|e^{t+1}| = |b - m^{t+1}(y^t) \odot y| = |b - \frac{b}{y^t} \odot y| = |\frac{b}{y^t} \odot (y - y^t)| \quad (15)$$

We finally obtain:

$$|e^{t+1}| = |\frac{b}{y^t}| |e^t(y)| \quad (16)$$

In this case, for $b < y^t$ (or $x^t > 0$), we obtain $|e^{t+1}| \leq |e^t(y)|$. Under these conditions the error will decrease for non-zero estimates of the unobserved signal.

Slowly Varying m^{t+1} : In the general case where NES is not locally invariant around y_t , let us assume m^{t+1} changes slowly enough so that there exists a constant λ satisfying:

$$|b - m^{t+1}(y) \odot y| \leq \lambda |b - m^{t+1}(y^t) \odot y| \quad (17)$$

It is possible to view λ as a measure of generalization. m^{t+1} satisfying better generalization properties will have a lower value of λ . Perfect generalization is recovered with $\lambda = 0$, and invariant m^{t+1} is achieved with $\lambda = 1$. For general λ , the error is at most larger than Eq. 16 by a factor of λ :

$$|e^{t+1}| \leq \lambda |\frac{b}{y^t}| |e^t| \quad (18)$$

We can immediately see that convergence will occur for elements satisfying: $|\frac{b}{y^t}| < \frac{1}{\lambda}$. For increasing values of λ i.e. decreasing generalization, only larger estimated values of x^t (relative to b) will achieve decreased error.

A good initialization of m^0 improves its generalization ability, decreasing its value of λ . The lower λ values increase the radius of convergence. This may explain the improved performance of better initializations.

6. Discussion

LMM vs. Adversarial Masking: LMM as a stand alone technique usually performed worse than Adversarial Masking, but served as a better initialization. We speculate that

mode collapse, inherent in adversarial training, makes the adversarial mask a lower bound on the \mathcal{X} source distribution. LMM can result in models that are too loose (i.e. also encode samples outside of \mathcal{X}). But as an initialization for NES, it is better to have a model that is too loose than a model which is too tight.

Automatic Label Extraction: To improve sample efficiency, we hypothesize that it would be possible to label only a limited set of examples as containing the target sound and not, and to use this seed dataset to train a deep sound classifier to extract more examples from an unlabeled dataset. We leave this investigation to future work.

Signal-Specific Losses: To showcase the generality of our method, we chose not to encode task specific constraints. In practical applications of our method however we believe that using signal-specific constraints can increase performance. Examples of such constraints include: repetitiveness of music (Rafii & Pardo, 2011), sparsity of singing voice, smoothness of natural images.

Additive and Convolution Mixtures: In line with most of the literature, our approach separates additive mixtures. In some settings, the mixtures are convolutional. We leave the expansion of NES to this setting for future work.

Non-Adversarial Alternatives: The good performance of LMM vs. AM on the vocals separation task, suggests that non-adversarial generative methods may be superior to adversarial methods for separation. This has also been observed in other mapping tasks e.g. the improved performance of NAM (Hoshen & Wolf, 2018) over DCGAN (Radford et al., 2015).

7. Conclusions

In this paper we proposed a novel method, Neural Egg Separation, for separating mixtures of observed and unobserved distributions. We showed that careful initialization using LMM improves results in challenging cases. Our method achieved much better performance than other methods and was usually competitive with full-supervision. Analytical results were presented to motivate the success of our method.

Acknowledgements

We thank Lior Wolf for fruitful discussions and for coining the name "Egg Separation".

References

- Barker, T. and Virtanen, T. Semi-supervised non-negative tensor factorisation of modulation spectrograms for monaural speech separation. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 3556–3561. IEEE, 2014.
- Bojanowski, P., Joulin, A., Lopez-Paz, D., and Szlam, A. Optimizing the latent space of generative networks. In *ICML'18*, 2018.
- Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *TASSP*, 1979.
- Chung, J. S. and Zisserman, A. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016.
- Davies, M. E. and James, C. J. Source separation using single channel ica. *Signal Processing*, 87(8):1819–1832, 2007.
- Denton, E. L. et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pp. 4414–4423, 2017.
- Gabbay, A., Ephrat, A., Halperin, T., and Peleg, S. Seeing through noise: Visually driven speaker separation and enhancement. *arXiv preprint arXiv:1708.06767*, 2017.
- Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT Press, 2016.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vaе: Learning basic visual concepts with a constrained variational framework. 2016.
- Hoshen, Y. and Wolf, L. Nam: Non-adversarial unsupervised domain mapping. In *ECCV'18*, 2018.
- Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5 (Nov):1457–1469, 2004.
- Huang, P.-S., Chen, S. D., Smaragdis, P., and Hasegawa-Johnson, M. Singing-voice separation from monaural recordings using robust principal component analysis. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 57–60. IEEE, 2012.
- Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. Deep learning for monaural speech separation. In *ICASSP*, 2014.
- Kim, H. and Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23 (12):1495–1502, 2007.
- Kim, J. and Park, H. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.
- Kim, T., Cha, M., Kim, H., Lee, J., and Kim, J. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010.
- Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.
- Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Smolley, S. P. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2813–2821. IEEE, 2017.
- Mimilakis, S. I., Drossos, K., Santos, J. F., Schuller, G., Virtanen, T., and Bengio, Y. Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask. *arXiv preprint arXiv:1711.01437*, 2017.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.
- Piczak, K. J. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018. ACM Press, 2015.
- Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- Rafii, Z. and Pardo, B. A simple music/voice separation method based on the extraction of the repeating musical structure. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 221–224. IEEE, 2011.
- Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R. The MUSDB18 corpus for music separation, December 2017. URL <https://doi.org/10.5281/zenodo.1117372>.

- Roweis, S. T. One microphone source separation. In *Advances in neural information processing systems*, pp. 793–799, 2001.
- Smaragdis, P., Raj, B., and Shashanka, M. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 414–421. Springer, 2007.
- Stoller, D., Ewert, S., and Dixon, S. Adversarial semi-supervised audio source separation applied to singing voice extraction. *arXiv preprint arXiv:1711.00048*, 2017.
- Stöter, F.-R., Liutkus, A., and Ito, N. The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 293–305. Springer, 2018.
- Subakan, C. and Smaragdis, P. Generative adversarial source separation. *arXiv preprint arXiv:1710.10779*, 2017.
- Vincent, E., Gribonval, R., and Fevotte, C. Performance measurement in blind audio source separation. *Trans. Audio, Speech and Lang. Proc.*, 14(4):1462–1469, 2006. ISSN 1558-7916.
- Wang, Y., Narayanan, A., and Wang, D. On training targets for supervised speech separation. *TASLP*, 2014.
- Wilson, K. W., Raj, B., Smaragdis, P., and Divakaran, A. Speech denoising using nonnegative matrix factorization with priors. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4029–4032. IEEE, 2008.
- Yu, A. and Grauman, K. Fine-grained visual comparisons with local learning. In *CVPR*, 2014.
- Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP*, 2017.
- Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pp. 597–613. Springer, 2016.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.