Separate And Diffuse: Using a Pretrained Diffusion Model for Improving Source Separation

Shahar Lutati[1]   **Eliya Nachmani**[2]   **Lior Wolf**[1]
[1]Department of Computer Science   [2]School of Electrical Engineering
Tel Aviv University
{shahar761,enk100}@gmail.com
wolf@mail.tau.ac.il

Abstract

The problem of speech separation, also known as the cocktail party problem, refers to the task of isolating a single speech signal from a mixture of speech signals. Previous work on source separation derived an upper bound for the source separation task in the domain of human speech. This bound is derived for deterministic models. Recent advancements in generative models challenge this bound. We show how the upper bound can be generalized to the case of random generative models. Applying a diffusion model Vocoder that was pretrained to model single-speaker voices on the output of a deterministic separation model leads to state-of-the-art separation results. It is shown that this requires one to combine the output of the separation model with that of the diffusion model. In our method, a linear combination is performed, in the frequency domain, using weights that are inferred by a learned model. We show state-of-the-art results on 2, 3, 5, 10, and 20 speakers on multiple benchmarks. In particular, for two speakers, our method is able to surpass what was previously considered the upper performance bound.

1   Introduction

Here is a simple algorithm for improving source separation. Given a test mixture m of C speakers, (1) apply a deep neural architecture B to m and obtain multiple approximated sources $\bar{v}_d^i$ for $i = 1...C$, (2) apply a generative diffusion model GM using each of the approximations obtaining $\bar{v}_g^i$, and (3) apply a shallow convolutional neural network F to $\bar{v}_d^i$ and $\bar{v}_g^i$ to obtain mixing weights $[\alpha_i, \beta_i] = F(\bar{v}_d^i, \bar{v}_g^i)$ and combine the two approximations linearly in the frequency domain to obtain the output $\bar{v}^i$.

The need to combine in the frequency domain arises because the reconstructed phase in each segment can be arbitrary and phase compensation is needed. In other words, since $\bar{v}_d$ and $\bar{v}_g$ are inferred by different processes, they may be at different phases.

In our experiments, out of the three networks (B, GM, F), we only train network F, as this training takes place on the same training set used to train network B. The other networks are taken, as is, from published models. Specifically, B is either Gated-LSTM [24] or SepFormer [30] and GM is the DiffWave network [14], which is trained, in an unsupervised way, on the LibriMix dataset [26] and WSJ0 [6] in order to generate speech signals.

Our empirical results, presented in Sec. 5, demonstrate that across multiple deep architectures and methods, applying a pretrained generative diffusion model, in the most straightforward way, pushes the envelope of results further by a significant gap. Our goal is to shed light on this phenomenon. We show the following two results: (1) the mutual information between the best combination of $\bar{v}_d^i$ and $\bar{v}_g^i$ and the underlying ground truth signal v^i is bounded by twice the mutual information between

the mixture and the ground truth signal, (2) we provide a bound for the signal-to-distortion ratio (the SDR error metric) of such combinations $\bar{v}^i$, which depends on the quality of network B and the mutual information between m and each source v^i.

2 Related work

Single-channel speech separation is a fundamental problem in speech and audio processing that has been extensively studied over the years [18, 23]. Recently, deep learning models have been proposed for speech separation, resulting in a significant improvement in performance compared to traditional methods. Hershey et al. proposed a clustering method that utilizes trained speech embeddings for separation [8]. Yu et al. proposed the Permutation Invariant Training (PIT) at the frame level for source separation [32], while Kolbaek et al. extended this approach by proposing the utterance-level Permutation Invariant Training (uPIT) [12]. An influential deep learning method for speech separation over the time domain was introduced by Luo et al. [20]. This method employs three components: an encoder, a separator, and a decoder. Subsequently, in Conv-Tasnet the separator network was replaced with a fully convolutional model, using a block of time-depth separable dilated convolution [21]. Conv-Tasnet was scaled by training several separator networks in parallel to perform an ensemble [34]. Dual Path RNN blocks were used to reorder the encoded representation and process it across different dimensions [19]. The so-called MulCat blocks were presented as a way to eliminate the need for a masking sub-network [24].

One of the limitations of current methods is their inability to effectively train neural networks for a large number of speakers, due to the reliance on Permutation Invariant Training (PIT) methods, which have a time complexity of $O(C!)$, where C is the number of speakers. Instead, one can use a permutation-invariant training method that employs the Hungarian algorithm, reducing the time complexity to $O(C^3)$ [5].

Lutati et al. introduced an upper bound for audio source separation [22]. By dividing the speech signal into short segments of sounds, a known distribution is used to describe the signal. Then, using the relation between mutual information and Cramer Rao lower bound, the authors managed to demonstrate an upper bound for speech separation for any deterministic model. This bound depends on the mutual information between the mixture and the sources. As the number of sources increases, the mutual information decreases, and so does the upper bound. The recent advent of very successful randomized generative models points in a new research direction: while the previous bound is applicable to any deterministic processing, it remains an open question whether it holds for generative models too.

The Diffusion Probabilistic Model has been successfully applied to various domains, such as time series and images [28]. A major limitation of this model is that it requires a significant number of iterative steps to generate valid data samples. This was addressed by a diffusion generative model based on Langevin dynamics and the score matching method [29]. This model estimates the Stein score function [17], which is the gradient of the logarithm of data density, and uses it to generate data points. Generative neural diffusion processes for speech generation were developed based on score matching [2, 14]. These models have achieved state-of-the-art results for speech generation, demonstrating superior performance compared to well-established methods, such as Wavernn [10], Wavenet [25], and GAN-TTS [1].

Jayaram et al. introduced the use of generative models as priors for the separation of different sources from a mixture [9]. This was tested on mixtures of images such as MNIST [16] and LSUN [33] datasets. DiffSep [27] performs audio separating using a diffusion model. The underlying SDE is an affine transformation of a mixing matrix P. While the authors showed the ability to separate sources from a single channel, their result falls behind deterministic models by a large margin. Although they set the ground for source separation using generative models, neither work presented state-of-the-art performance, nor proposed a bound for generative methods.

We are unaware of similar generalization bounds developed for diffusion models. Our method can be applied to other domains, such as image denoising and conditional generation, given suitable data on the underlying distribution of the training data. For example, for image denoising, there are known results from the field of natural image statistics [31] that state that while segmenting the image into patches a Gaussian Scale Mixture is able to model the statistics of the underlying natural images. A

similar derivation to what is done in our work for audio would obtain an upper bound for the maximal improvement achievable given the noise prior.

3 Analysis

Let m be a mixture of C speakers, each with a ground truth signal v^i, $i = 1 \dots C$. Let B be a source separation model that returns C signals and GM be a vocoder diffusion model that is trained on the domain of clear, single-speaker voice signals. The method described at the beginning of Sec. 1 can be summarized by the following set of equations, where STFT and iSTFT are the Short-Time Fourier transform and its inverse, respectively.

$$[\bar{v}_d^1, \bar{v}_d^2, \dots, \bar{v}_d^c] = B(m) \tag{1}$$

$$\forall i \in [c] \begin{cases} \bar{v}_g^i = GM(\bar{v}_d^i) & (2) \\ \bar{V}_g^i, \bar{V}_d^i = \text{STFT}(\bar{v}_g^i), \text{STFT}(\bar{v}_d^i) & (3) \\ [\alpha_i, \beta_i] = F(\bar{V}_d^i, \bar{V}_g^i) & (4) \\ \bar{v}^i = \text{iSTFT}(\alpha_i \odot \bar{V}_d^i + \beta_i \odot \bar{V}_g^i) & (5) \end{cases}$$

Eq. 1 applies B to separate the mixed signal. Eq. 2 applies the vocoder to denoise each output of B separately. Eq. 4 computes the combination weights of the output of B and the corresponding output of GM. Finally, the combination is performed by linearly mixing in the spectral domain in Eq. 5. Where $\odot$ is the Hadamard product.

As we show in Sec. 5, this simple method empirically surpasses the current state-of-the-art. Evidently, when applied correctly, a vocoder model can suppress the errors that the deterministic source separation model has. Below, we analyze the source of this improvement and derive an upper bound for the level of improvement.

Definition 3.1 (Signal-to-Distortion Ratio (SDR)). Let the error ϵ be defined as the difference between the source and the estimated source. The Signal-to-Distortion Ratio (SDR) is $SDR = 10log_{10}\frac{Var(v)}{Var(\epsilon)}$.

Given a mixture of sources, m, the ground truth signal v, and the estimated signal, $\bar{v}_d$, [22] have found that

$$SDR(v, \bar{v}_d) \leq 10log_{10}\left(\frac{L}{w} \cdot Var(v) \cdot I(m_r, v_r)\right) \tag{6}$$

where $\bar{v}_d$ is the estimation of the deterministic backbone network B, L is the length of the signal, w is the segment width, v is the ground truth source, m is the mixture and the subscript r describes the mixture and the ground truth signal in the r-th segment.

Furthermore, [22] show that the mutual information between the deterministic estimation and the source is upper bounded by the mutual information between the source and the mixture.

$$I(v_r; \bar{v}_{dr}) \leq I(v_r; m_r) \tag{7}$$

In what follows, we first lay down the basis for upper bounding the combination of the generative and the deterministic signal. Next, by modeling the noise that is added to the generated signal, the mutual information between the combination of the signals and the sources is bounded. Finally, by applying the changes to Eq. 6, a new upper bound is found.

Lemma 3.2 (Mutual Information Chain Rule). *For any random variables X,Y,Z the following chain rule holds $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$.*

See proof in the supplementary. Using Lemma. 3.2 for the source, v_r, the deterministic estimation, $\bar{v}_{dr}$, and the generative estimation, $\bar{v}_{gr}$, we have,

$$I(v_r; \bar{v}_{dr}, \bar{v}_{gr}) = I(v_r; \bar{v}_{dr}) + I(v_r; \bar{v}_{gr}|\bar{v}_{dr}) \tag{8}$$

Plug Eq. 8 to Eq. 7, we have the following bound

$$I(v_r; \bar{v}_{dr}, \bar{v}_{gr}) \leq I(v_r; m_r) + I(v_r; \bar{v}_{gr}|\bar{v}_{dr}) \tag{9}$$

Our goal is now to bound the maximal mutual information achievable between the source, v_r, and the generative signal, $\bar{v}_{gr}$.

3

3.1 Generative Bound

The inherent noise of GM is modeled as additive noise. For any practical GM there is an inherent error in reconstructing a perfect signal from a perfect prior.

The pretrained generative model (GM) is optimized to generate samples v_{gr} that minimize the ELBO loss when given a prior for v_r.

Denote the observed prior as $p(v_d r)$. GM generates v_{gr} that it is the optimizer for ELBO loss:

$$p(v_{gr}) = argmin_{p(v_{gr})} ELBO(p(v_{dr}), p(v_{gr})) \tag{10}$$

Given the aforementioned upper bound $I(v_r; v_{dr})$, Eq. 7, it follows that $p(v_{gr}) \approx p(v_{dr})$ when $I(v_r; v_{dr}) \leq I(v_r; m)$. This implies that the v_{gr}, will sample from the same distribution as v_{dr}. Using Eq. 9 We have, $I(v_r; v_{gr})$ is upper bounded by $I(v_r; m_r)$, and $I(v_r; v_{gr}, v_{dr})$.

$$I(v_r; v_{gr}) \leq I(v_r; m_r) + I(v_r; v_{gr}, v_{dr}) \leq 2 \cdot I(v_r; m_r) \tag{11}$$

Let us assume the added noise is Gaussian noise. From [4], we have that among all distributions Gaussian noise is the worst-case additive noise in terms of mutual information, that is, normal additive noise will degrade most of the information in the generated signal. As shown below, for the pretrained GM, the worst case coalesces with the best case, where the inequality in Eq. 11 becomes equality. Thus, making Eq. 11 tight.

For additive white Gaussian noise (AWGN) it holds that $\bar{v}_{gr} = x + n, n \sim \mathcal{N}$ where x is the desired part of the signal, based on the prior, and n is the normal noise.

$$Pr(n) = \frac{exp(-\frac{n^2}{2\sigma^2})}{\sqrt{2\pi\sigma^2}} \tag{12}$$

Where σ^2 is the variance of the additive noise. [22] show empirically that for short segments of 20[ms] the audio is distributed in the time domain as a Laplace distribution, $x \sim \mathcal{L}aplace$.

$$Pr(x) = \frac{exp(-|\frac{x}{\sqrt{2}}|)}{\sqrt{2}} \tag{13}$$

Computing explicitly the mutual information we obtain

$$I(x; \bar{v}_g) = \int_x \int_{\bar{v}_g} Pr(x, \bar{v}_g) \log \left(\frac{Pr(x, \bar{v}_g)}{Pr(x) Pr(\bar{v}_g)} \right) dx \, d\bar{v}_g, \tag{14}$$

where $Pr(x)$ is the probability of x, $Pr(\bar{v}_g)$ is the probability of the generated voice, and $Pr(x, \bar{v}_g)$ is the joint probability of the signals.
The joint probability can be written explicitly as,

$$Pr(x, \bar{v}_g) = Pr(\bar{v}_g|x)Pr(x) \tag{15}$$

Plugging in the known distributions reads

$$Pr(x, \bar{v}_g) = \frac{exp(-\frac{1}{2}(\frac{\bar{v}_g - x}{\sigma})^2)}{\sqrt{2\pi\sigma^2}} \cdot \frac{exp(-|\frac{x}{\sqrt{2}}|)}{\sqrt{2}} \tag{16}$$

The distribution of $\bar{v}_g$ can be formulated using marginalization of the joint probability, $Pr(x, \bar{v}_g)$

$$Pr(\bar{v}_g) = \int_x Pr(x, \bar{v}_g) \, dx \tag{17}$$

Recall that the generated voice without noise can at best (maximum) obtain the mutual information between the mixture and the sources. Let ρ be the ratio between the mutual information of the source and the generative estimation, and the mutual information of the source and mixture, i.e., $\rho = \frac{I(v_r, \bar{v}_g)}{I(v_r, m_r)}$. The factor ρ is computed numerically from equation Eq. 14, by integrating the double integral.
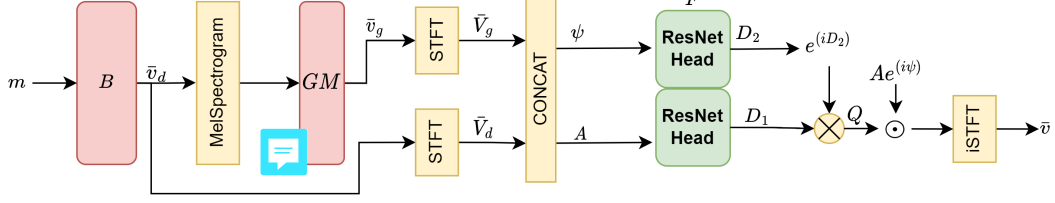
Figure 2: The suggested system architecture. Red blocks indicate pretrained models, Yellow blocks indicate non-learned operations, and Green blocks indicate learned neural models.

Figure 1 depicts the ratio ρ as defined above. This ratio is computed for numerous sources. Each trial is with a different color. Note that for low-quality generative models (i.e, $\sigma^2 \geq 10^{-2}$) the ratio degrades at a faster rate when the number of sources is lower. Recall that when the number of sources is larger, the mixture tends to have a normal distribution. Thus, the addition of another independent noise with a smaller variance has a negligible effect in terms of the mutual information (which is already low). Observe that ρ is approximately one when $\sigma^2 \leq 10^{-2}$. Since, for any well-functioning pretrained GM, the error in reconstructing the audio signal is below 10^{-2} [14, 2, 3], we obtain that

Figure 1: ρ as a function of the noise variance.

$$\rho(\sigma^2 \leq 10^{-2}) \approx 1 \quad \forall C \in [2, 20] \tag{18}$$

Therefore, rewriting Eq. 11 with the definition of ρ reads,

$$\rho I(v_r; m_r) = I(v_r; \bar{v}_g) \leq I(v_r; \bar{v}_{dr}) \leq I(v_r; m_r). \tag{19}$$

Using the result of Eq. 18 one obtains,

$$I(v_r; m_r) \approx I(v_r; \bar{v}_g) \leq I(v_r; \bar{v}_{dr}) \leq I(v_r; m_r), \tag{20}$$

which implies that the bound is tight. Also, from the bound, under the reasonable assumptions made, the deterministic approximation and the generative approximations would have similar error magnitudes.

Finally, updating Eq. 6 with the new bound obtains,

$$SDR(v, \bar{v}) \leq 10 log_{10} \left(\frac{L}{w} \cdot Var(v) \cdot I(m_r, v_r) \right) + 3.0 \tag{21}$$

One can see a maximum addition of 3dB to the upper bound when combining both generative results with deterministic processing. Note that this bound is tight in the sense that the worst-case additive independent noise is taken into account and still the achievable mutual information is at maximum.

4   Method

Given a mixture signal m, the deterministic estimation $\bar{v}_d$ is obtained using the backbone network B. Then, a Mel-Spectrogram is computed over all $\bar{v}_d$ estimations.

$$Mel(\bar{v}_d) = MelSpectrogram(\bar{v}_d) \tag{22}$$

Using a pretrained vocoder, GM, specifically DiffWave, and the priors obtained from the deterministic backbone B, a generative estimation is obtained $\bar{v}_g$.
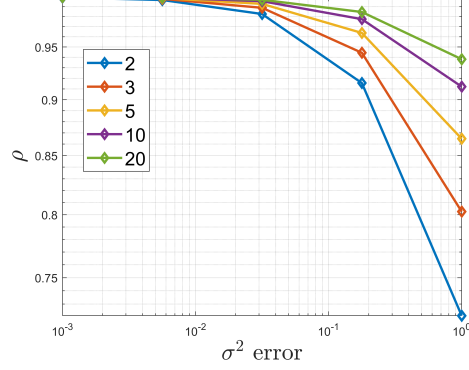
$$\bar{v}_g = GM(Mel(\bar{v}_d)) \tag{23}$$

5

Since the generative vocoder is not given any phase information, the generated signal has a phase shift that can vary over time. To combine both estimations, an alignment procedure is employed.

In order to align the signals, both estimations are transformed into the frequency domain, where the aligning operation dual is multiplication by a phasor. The transformation is done through a short-time Fourier transform with N_{FFT} frequency bins and K segments.

Denote the spectrogram of $\bar{v}_d^i$ and $\bar{v}_g^i$, as $\bar{V}_d^i, \bar{V}_g^i$ respectively, $\bar{V}_d^i, \bar{V}_g^i \in \mathbb{C}^{N_{\text{FFT}} \times K}$. The absolute phase of $\bar{V}_g$ is of no importance; what is important is its trend over short segments. The phase of $\bar{V}_d$ and the relative phase between $\bar{V}_d$ and $\bar{V}_g$ is concatenated into a 2-channel tensor, $\psi \in \mathbb{R}^{C \times 2 \times N_{\text{FFT}} \times K}$, i.e.,

$$\psi = Concat(\angle \bar{V}_d, \angle(\bar{V}_g \odot \bar{V}_d^*)), \tag{24}$$

where $\odot$ is the Hadamard product, the star superscript denotes conjugate, and the following definition is used.

Definition 4.1. The angle, denoted as $\angle$, of a complex number is computed by

$$\angle X = tan^{-1}\left(\frac{Im(X)}{Re(X)}\right) \tag{25}$$

where $X \in \mathbb{C}$, Im is the imaginary part, and Re is the real part.

In addition to the relative phase tensor ψ, the magnitude of both $\bar{V}_d^i$ and $\bar{V}_g^i$ is concatenated into tensor, A,

$$A = Concat(|\bar{V}_d|, |\bar{V}_g|) \tag{26}$$

The alignment network that we train F is a dual 6-layer convolutional neural network, with residual connections [7], termed ResNet Head in Fig. 2. The magnitude and phase heads share the same hyperparameters. The processed tensors are combined into the complex factor $Q \in \mathbb{C}^{C \times 2 \times N_{FFT} \times K}$,

$$D_1, D_2 = F(A, \psi) \tag{27}$$

where D_1 and D_2, both in $\mathbb{R}^{C \times 2 \times N_{FFT} \times K}$, represent the magnitude and phase, respectively, of a complex tensor Q. This dual representation is used as an effective representation for complex tensors. Q is then computed explicitly by combining the magnitude and phase

$$Q = D_1 \cdot exp(-jD_2) \tag{28}$$

Below, we denote the different channels of Q by a subscript i. Define as $\alpha = Q_1 \in \mathbb{C}^{C \times 1 \times N_{FFT} \times K}$ and the second channel as $\beta = Q_2$.

α and β are the coefficients used in Eq. 5. Written more explicitly, the weighted sum of $\bar{V}_d$ and $\bar{V}_g$ is computed as

$$\bar{V} = \alpha \odot \bar{V}_d + \beta \odot \bar{V}_g \tag{29}$$

Finally, the time-domain signal is obtained by employing the inverse short-time Fourier transform.

$$\bar{v} = iSTFT(\bar{V}) \tag{30}$$

Objective Function   Following [21, 30, 22, 27], the objective function is the Scale-Invariant SDR. The scale-invariant SDR is agnostic to the scale of the estimated signal, and also to spurious errors, as presented in [15]. First, project the source onto the estimated signal.

$$\tilde{v} = \frac{<v, \bar{v}> v}{||v||^2} \tag{31}$$

Second, compute the normalized error

$$\tilde{e} = \bar{v} - \tilde{v} \tag{32}$$

Third, the scale-invariant SDR is computed as follows,

$$SI - SDR(v, \bar{v}) = 10log10\left(\frac{||\tilde{v}||^2}{||\tilde{e}||^2}\right) \tag{33}$$

For a large number of speakers ($C \geq 10$), the Hungarian algorithm obtains better results than the permutation invariant loss, by explicitly assigning the different estimated sources to the ground truth sources [5]. Therefore, network F is trained using the assignment obtained by the Hungarian method for the deterministic voice separation network.

6

5 Experiments

For all datasets, the same alignment network architecture is employed, with the same hyper-parameters. It is a CNN with six layers, 3×3 kernels, and 32,32,64,64,64 hidden channels. For Librimix and WSJ0 datasets the DiffWave was trained separately over the training sets' sources (no mixing). The pretrained models for separation are taken from the official publication when available and from HuggingFace hub otherwise. The optimization procedure is done with Adam [11] optimizer with a learning rate of 1E-3 and a batch size of 3. The setting involved 3 GPUs each Nvidia A5000.

Datasets   The LibriSpeech dataset [26] is a large corpus of read English speech, designed for training and evaluating automatic speech recognition systems. It consists of over 1000 hours of audio from audiobooks read by professional and non-professional speakers.

The Wall Street Journal (WSJ0) [6] dataset is a collection of read English speech samples and is widely used by the research community. The dataset includes a total of 80 hours of audio.

Following previous work [30, 5], mixtures are generated by a random process. The speakers are divided into training speakers and test speakers. A training- or a test sample is created by combining random speakers out of the respective set of speakers, with random SNR values between $0 - 5$ dB.

Baseline methods   We compare our method to the state-of-the-art methods in voice separation, including DiffSep [27] that employs diffusion models, SepFormer [30], which is a transformer model, and SepIt [22], which extends the Gated-LSTM method [24] by running it in an iterative manner.

As a generative model in our experiments, we employ a pretrained DiffWave model [14].

In addition to running our method, we perform an ablation that is aimed at verifying the need for a learned combination model F. In this ablation, we align (recover the phase) the generative approximation with the deterministic one and then average the two. Another ablation is using the GAN-based model HiFiGAN [13], trained on WSJ0, which is a deterministic generative model (no noise is given to the generator). Since this generative model is deterministic the previous upper bound applies and we verify experimentally that no significant improvement is obtained.

First, the cross-correlation function is used as a measure of similarity. In the frequency domain, the cross-correlation dual is computed using conjugate multiplication.

$$\bar{R} = \bar{V}_d \odot \bar{V}_g^* , \tag{34}$$

where $\bar{R}$ is the frequency domain. For each segment, the inverse Fourier transform is conducted, and then the time for which the signal is at maximum is searched.

$$t_i^* = \arg\max(IFFT(\bar{R}[I]) \quad \forall i \in \{1, \dots, K\} , \tag{35}$$

where t_i^* is the time delay within the i-th segment, $\bar{R}[I]$ is the i-th frequency domain cross correlation segment.

A complex phase shift corresponding to the time delay found per segment is computed via the following,

$$T_i = exp(j(2\pi f t_i^*)) \quad \forall i \in \{1, \dots, K\} , \tag{36}$$

where T_i is the complex Fourier factor of time delay, and f is the frequency that matches each frequency bin in the spectral transform.

The combination of the two estimations now reads,

$$\bar{V}_{xcorr} = \bar{V}_d + T \odot \bar{V}_g \tag{37}$$

The time domain signal is obtained as before, using the inverse short-time Fourier transform.

$$\bar{v}_{xcorr} = iSTFT(\bar{V}_{xcorr}) \tag{38}$$

Qualitative Results   A visualization of $\bar{v}_d$ and $\bar{v}_g$ is depicted in Fig. 3. Evidently, both signals have a phase difference, which changes over time. For example, in panel (b), at [0.52,0.56] the phase of the generative estimation tracks the phase of the deterministic model. At the other end, at [0.56,0.64], the phase between the signal is offset by π. Recall that while $\bar{v}_d$ is given the full information about the
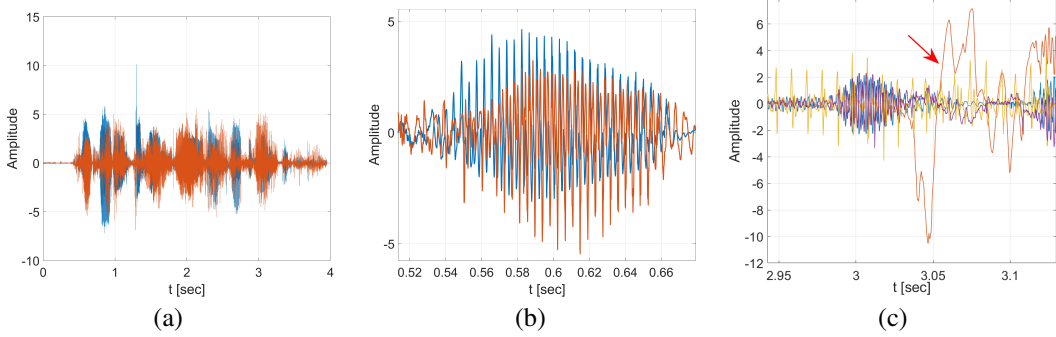
Figure 3: (a) Visualization of $\bar{v}_d$, and $\bar{v}_g$, (b) A zoom-in that shows the phase difference between two estimations. Blue-$\bar{v}_d$, Orange-$\bar{v}_g$. (c) An example where $\bar{v}_g$ is more precise than $\bar{v}_d$, and the resulting combination that tracks the source v. Blue-v, Orange-$\bar{v}_d$, Yellow-$\bar{v}_g$, Purple-$\bar{v}$.
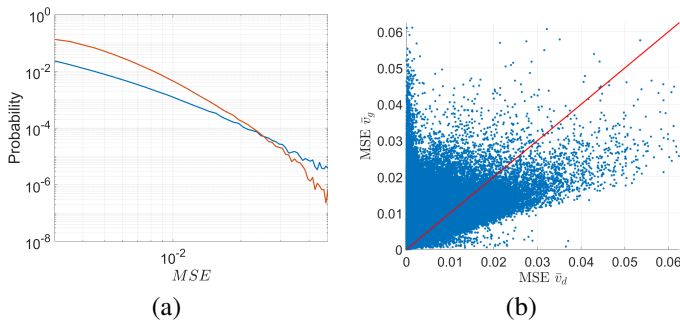

Figure 4: Comparing the error rates of the deterministic and the generative parts. (a) The histograms of the Mean Square Error (MSE) between v to $\bar{v}_d$ (Blue), and v to $\bar{v}_g$ (Orange). (b) Scatter plot of the MSE for $\bar{v}_d$ (x-axis) vs. $\bar{v}_g$ (y-axis).

phase, $\bar{v}_g$ is given only magnitude information and thus the absolute phase is not synched. Therefore, the alignment procedure is necessary.

The generated signal, $\bar{v}_g$, does not track the deterministic part, $\bar{v}_d$, one-to-one, but sometimes fixes sudden errors of other sources that the deterministic model could not handle correctly. This phenomenon is depicted in Fig. 3(c), where the deterministic signal has some bursts. As can be seen, the combined signal is more related to the ground truth signal.

The supplementary presents samples of both the deterministic and the generative estimation. While the deterministic estimation presents less background noise, it misses multiple temporal parts, which are filtered out by the consecutive chain of static filters of the deterministic model. The signal produced by the generative model, on the other hand, maintains a considerable amount of background noise but is able to sample parts that were previously omitted. The combination of both estimations yields a signal that is closer to the desired source.

To estimate the relative quality of the deterministic and generative estimations, the Mean Square Error between the different estimations and the aligned sources is calculated in short segments of 20[ms]. The histogram depicted in Fig. 4(a) shows that the generative signal $\bar{v}_g$ has a slight tendency to greater error, but has the same performance overall. This is in agreement with Eq. 20, where we expect that both $\bar{v}_g$, and $\bar{v}_d$ obtain similar performances. This result confirms that a combination of both estimations should improve the overall estimation of sources. A scatter plot of the errors is depicted in Fig. 4(b). The red line indicates a ratio of one to one between the deterministic error and the generative one. As can be seen, the error of the generative estimation is within the same range as that of the deterministic estimation, and there is no estimation that is markedly better than the other.

Source separation results The voice separation results for all the benchmarks are depicted in Tab. 1. For the WSJ0 dataset, when separating a mixture of two sources, the improvement in terms of SI-SDR is 1.5dB over the current state-of-the-art. Using our method not only improved the SI-SDR beyond all other methods, but also broke the previous bound for deterministic models. This emphasizes the strength of the diffusion models as holding independent modeling benefits that can be used to improve speech separation. For three sources, separation is improved by 1dB, again obtaining state-of-the-art results, but somewhat lower than the classical bound of deterministic methods. We

Table 1: The performance obtained by our method on the various WSJ0 and LibriSpeech benchmarks. The reported values are SI-SDRi $[dB]$ (higher is better). -F refers to cross-correlation ablation, where the mixing network F is not used. *These missing runs of applying our method to baseline models are due to the inability to obtain the pre-trained baseline models.

Method	WSJ0		LibriSpeech			
	2Mix	3Mix	2Mix	5Mix	10Mix	20Mix
Classical Upper Bound (Lutati et al.)	23.1	21.2	23.1	14.5	12.0	8.0
Generative Upper Bound (ours)	26.1	24.2	26.1	17.5	15.0	11.0
DiffSep [27]	14.3	-	-	-	-	-
SepIt [22]	22.4	20.1	-	13.7	8.2	-
SepFormer [30]	22.3	19.8	20.6	-	-	-
SepFormer + HiFiGAN [13] (ablation)	22.3	20.0	-	-	-	-
SepFormer + DiffWave -F (ablation)	22.6	20.3	20.8	-	-	-
SepFormer + DiffWave (ours)	**23.9**	**20.9**	**21.5**	-	-	-
Gated LSTM [24]	20.1	16.9	-	12.7	7.7	4.3
Gated LSTM + DiffWave -F (ablation)	-*	-*	-	13.0	8.1	4.5
Gated LSTM + DiffWave (ours)	-*	-*	-	**14.2**	**9.0**	**5.2**

note that the ablation that removes network F and introduces a heuristic alignment and equal mixing performs considerably better than the deterministic part (the SepFormer baseline), but only slightly better than the iterative SepIt. The deterministic HiFiGAN ablation shows little to no improvement, supporting the divide we make between deterministic and nondeterministic models.

For LibriSpeech we use an available SepFormer model as a baseline in the case of two speakers, and for the other mixtures, an available Gated LSTM model. As can be seen, for all numbers of sources, a major improvement is obtained over the current state of the art. The alignment ablation degrades the separation result and is inferior to the learned method, but is still competitive. Evidently, there is still room for improving separation methods. For five speakers, the result obtained falls 0.3dB short of surpassing the classical upper bound. For other mixtures, the gap is larger.

6 Limitations

The new bound we propose applies to any possible non-deterministic model. A tighter bound can be obtained, for example, in a bound that is developed specifically for diffusion models. Its non-specific nature also means that it does not provide practical guidelines as to the development of specific methods, with the exception that it encourages the addition of non-deterministic post-processing by pointing out the limited nature of the existing bound. The bound is developed, similar to the deterministic bound, in the context of algorithms that work on chunks of a limited size. As the community focuses on efficient attention models for large contexts, new methods can emerge for which the bound would not hold.

7 Conclusions

A general upper bound for source separation using generative models is proposed. The generalization of the upper bound suggests that for a pretrained generative model, a maximal improvement of 3dB from the previous deterministic bound is achievable. In addition, a simple yet effective method is suggested, where the deterministic signals and the generated signals are combined in the frequency domain. The combination procedure is learned and shown to be superior to the classical alignment method. When tested on various numbers of speakers, the estimation always yields improvement. For two sources, the separation is able to surpass the previous upper bound, demonstrating the need to go beyond a deterministic model.

The upper bound derived here is general in the sense that a similar bound would hold when the underlying distributions are replaced to fit other data domains. This is left for future work.

8 Acknowledgments

References

[1] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. High fidelity speech synthesis with adversarial networks. In *International Conference on Learning Representations*, 2019.

[2] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2020.

[3] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *arXiv preprint arXiv:2106.09660*, 2021.

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[5] Shaked Dovrat, Eliya Nachmani, and Lior Wolf. Many-speakers single channel speech separation with optimal permutation training. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

[6] Garofolo, John S., Graff, David, Paul, Doug, and Pallett, David. CSR-I (WSJ0) complete, 1993.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[9] Vivek Jayaram and John Thickstun. Source separation with deep generative priors. In *International Conference on Machine Learning*, pages 4724–4735. PMLR, 2020.

[10] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2018.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[12] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913, 2017.

[13] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[14] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2020.

[15] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. SDR–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[16] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[17] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

[18] G Logeshwari and GS Anandha Mala. A survey on single channel speech separation. In *International Conference on Advances in Communication, Network, and Computing*, pages 387–393. Springer, 2012.

[19] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. *arXiv preprint arXiv:1910.06379*, 2019.

[20] Yi Luo and Nima Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2018.

[21] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019.

[22] Shahar Lutati, Eliya Nachmani, and Lior Wolf. Sepit: Approaching a single channel speech separation bound. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association*, pages 5323–5327. ISCA, 2022.

[23] Rainer Martin and Israel Cohen. Single-channel speech presence probability estimation and noise tracking. *Audio Source Separation and Speech Enhancement*, pages 87–106, 2018.

[24] Eliya Nachmani, Yossi Adi, and Lior Wolf. Voice separation with an unknown number of multiple speakers. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7164–7175, 2020.

[25] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[26] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[27] Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaeuk Byun, Soyeon Choe, and Min-Seok Choi. Diffusion-based generative speech source separation. *arXiv preprint arXiv:2210.17327*, 2022.

[28] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[29] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[30] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE, 2021.

[31] Yair Weiss and William T Freeman. What makes a good model of natural images? In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[32] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017.

[33] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[34] Liwen Zhang, Ziqiang Shi, Jiqing Han, Anyan Shi, and Ding Ma. Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks. In *International Conference on Multimedia Modeling*, pages 653–665. Springer, 2020.

A Proof of Lemma 3.2

The proof of Lemma 3.2 in the main text is below.

Proof. Using the definition of mutual information, and chain rule for entropy

$$
\begin{aligned}
I(X;Y,Z) &= H(Y;Z) - H(Y;Z|X) \\
&= H(Z) + H(Y|Z) - H(Z|X) - H(Y|X,Z) \\
&= (H(Z) - H(Z|X)) + (H(Y|Z) - H(Y|X,Z)) \\
&= I(X;Z) + I(X;Y|Z)
\end{aligned}
\tag{39}
$$

$\square$

B Sample spectrograms

Figure 5 presents sample spectrograms. Shown are the source, the deterministic estimation, the estimation obtained by the generative model, and their combination in the frequency domain using the coefficients produced by the function F.
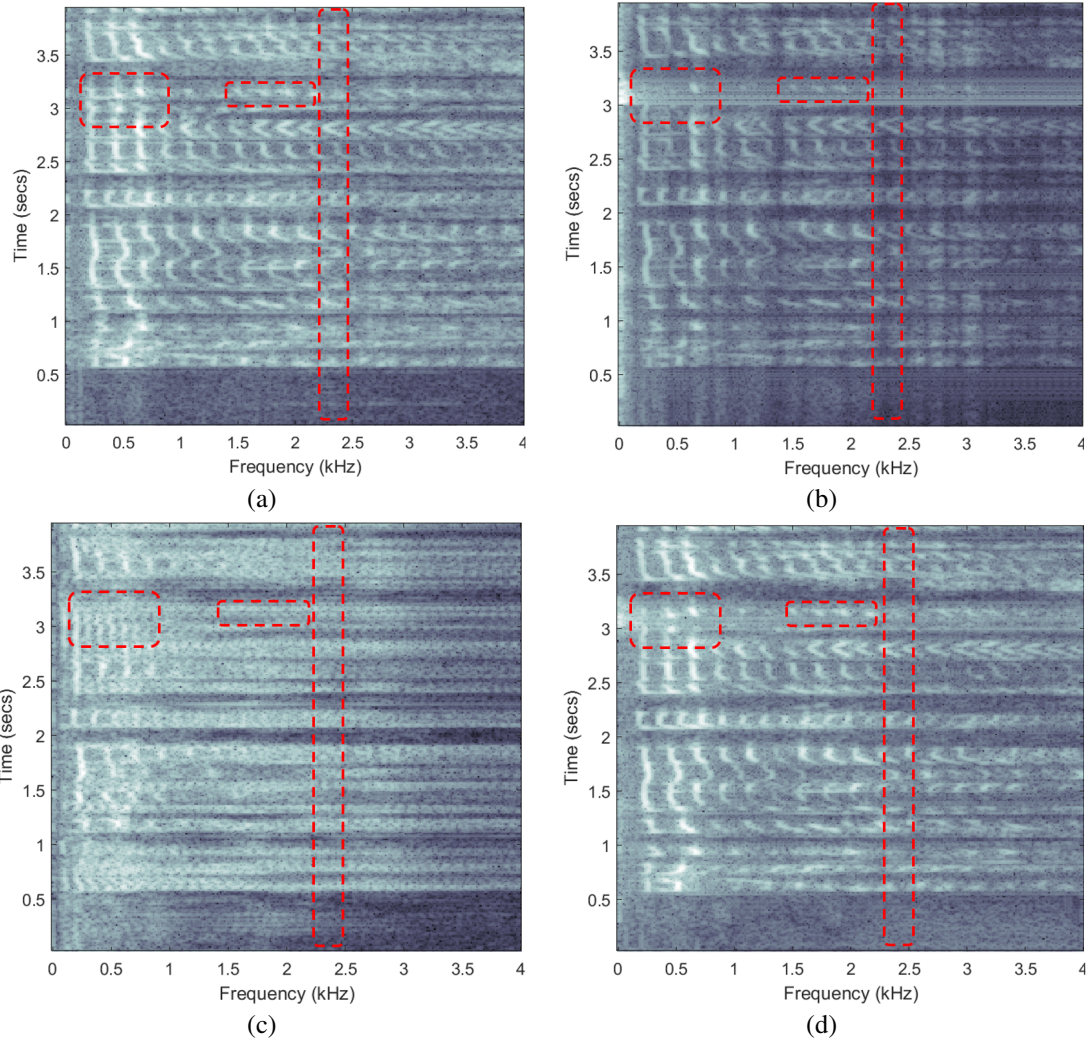
Figure 5: (a) Spectrogram of a source. (b) Spectrogram of the deterministic estimation $\bar{v}_d$ (c) Spectrogram of the generated voice $\bar{v}_g$ (d) Spectrogram of the combined estimation $\bar{v}$.