Interspeech 2018
2-6 September 2018, Hyderabad

Monaural Audio Source Separation using Variational Autoencoders

Laxmi Pandey[*1], *Anurendra Kumar*[*1], *Vinay Namboodiri*[1]

[1]Indian Institute of Technology Kanpur

laxmip@iitk.ac.in, anu.ankesh@gmail.com, vinaypn@iitk.ac.in

Abstract

We introduce a monaural audio source separation framework using a latent generative model. Traditionally, discriminative training for source separation is proposed using deep neural networks or non-negative matrix factorization. In this paper, we propose a principled generative approach using variational autoencoders (VAE) for audio source separation. VAE computes efficient Bayesian inference which leads to a continuous latent representation of the input data(spectrogram). It contains a probabilistic encoder which projects an input data to latent space and a probabilistic decoder which projects data from latent space back to input space. This allows us to learn a robust latent representation of sources corrupted with noise and other sources. The latent representation is then fed to the decoder to yield the separated source. Both encoder and decoder are implemented via multilayer perceptron (MLP). In contrast to prevalent techniques, we argue that VAE is a more principled approach to source separation. Experimentally, we find that the proposed framework yields reasonable improvements when compared to baseline methods available in the literature i.e. DNN and RNN with different masking functions and autoencoders. We show that our method performs better than best of the relevant methods with ~ 2 dB improvement in the source to distortion ratio.

Index Terms - Autoencoder, Variational inference, Latent variable, Source separation, Generative models, Deep learning

[*]

1. Introduction

The objective of Monaural Audio Source Separation (MASS) is to extract independent audio sources from an audio mixture in a single channel. Source separation is a classic problem and has wide applications in automatic speech recognition, biomedical imaging, and music editing. The problem is very challenging since it's an ill-posed problem i.e. there can be many combinations of solutions and the objective is to estimate the best possible solution. Traditionally, the problem has been well addressed by non-negative matrix factorization(NMF) [1] and PLCA[2] . These models learn the *latent bases* which are specific to a source from clean training data. These latent bases are later utilized for separating source from the mixture signal [3]. NMF and PLCA are generative models which work under the assumption that the data can be represented as the linear composition of low-rank latent bases. Several extensions of NMF and LVM have been employed in literature along with temporal, sparseness constraints [4, 1, 5]. Though NMF and PLCA are scalable, these techniques do not learn discriminative bases and therefore yield worse results when compared to models where bases are learned on mixtures. Discriminative NMF [6] has been proposed in order to learn mixture specific bases which in turn has shown some improvement over the NMF. NMF based approaches assume that data is a linear combination of latent bases and it may be a limiting factor for real-world data. To model the non-linearity, deep neural networks(DNN), in various different configurations have been used in source separation [7, 8, 9]. The denoising auto-encoder (DAE) is a special type of fully connected feedforward neural networks which can efficiently de-noise a signal [10]. They are used to learn robust low-dimensional features even when the inputs are perturbed with some noise [11] . DAEs have been used for source separation with input as a mixed signal and the output as the target source, both in form of spectral frames [12]. Though DAEs have a lot of advantages, it comes with the cost of high complexity and the loss in spatial information. Fully connected DAEs cannot capture the 2D (spectral-temporal) structures of the spectrogram of the input and output signals and have a lot of parameters to be optimized and hence the system is highly complex. The fully convolutional denoising autoencoders [13] maps the distorted speech signal to its clean speech signal with an application to speech enhancement. Recently, a deep (stacked) fully convolutional DAEs (CDAEs) is used for the audio single channel source separation (SCSS) [14]. However, current deep learning approaches for source separation are still computationally expensive with a lot of parameters to tune and not scalable. NMF based approaches, on the other hand, work with the simplistic assumption of linearity and the inability to learn discriminative bases effectively.

In this paper, our goal is to have best of both worlds - i) To learn a set of bases effectively (which is done by encoder and decoder in VAE) and ii) Inexpensive computation. Moreover, unlike other methods, VAE can also yield the confidence scores of how good or bad are the separated sources, based on the average posterior variance estimates. VAE has shown state-of-the-art in image generation, text generation and reinforcement learning [15, 16, 17, 18, 19]. In this paper, we show the effectiveness of VAE for audio source separation. We compare the performance of VAE with DNN/RNN architectures and autoencoders. VAE performs better than all methods in terms of a source to distortion ratio (SDR) with ~ 2 dB improvement.

2. Variational Autoencoder

The variational autoencoder [15] is a generative model which assumes that an observed variable x is generated from an underlying random process with latent variable z as random variables. In this paper, we aim to learn a robust latent representation of a noisy signal i.e. $P(z|x) \approx P(z|x+n)$, where x and n denotes signal and noise respectively. While estimating z for a source, we consider other sources as noise. The latent variable z is further used to estimate the clean (separated) source. Fig. 1 shows the graphical model of VAE.

[*]The first two authors contributed equally

10.21437/Interspeech.2018-1140

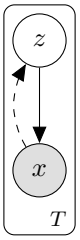


Figure 1: *Graphical model of VAE. T is total number of spectral frames. Dotted line denotes the inference of latent variable while solid line denotes the generative model of observed variable.*

Mathematically, the model can be represented as:

$$P_\theta(x, z) = P_\theta(x|z)P_\theta(z) \tag{1}$$

$$P_\theta(x) = \int P_\theta(x|z)P_\theta(z)dz \tag{2}$$

VAE assumes that the likelihood function $P_\theta(x|z)$ and prior distribution $P_\theta(z)$ come from a parametric family of distributions with parameters θ. The prior distribution is assumed to be a Gaussian with zero mean and unit variance:

$$P(z) = \mathcal{N}(z; 0, I) \tag{3}$$

The likelihood, is often modeled using an independent Gaussian distribution whose parameters are dependent on z,

$$P_\theta(x|z) = \mathcal{N}(x; \mu_\theta(z), \sigma^2_\theta(z)I) \tag{4}$$

where, $\mu_\theta(z)$ and $\sigma^2_\theta(z)$ are non-linear functions of z which is modeled using a neural network. The posterior distribution $P_\theta(z|x)$ can be written by Bayes's formula,

$$P_\theta(z|x) = \frac{P_\theta(x|z)P(z)}{\int P_\theta(x, z)dz} \tag{5}$$

However, the denominator is often intractable. Sampling methods like MCMC can be employed, but these are often too slow and computationally expensive. Variational Bayesian methods solves this problem by approximating the intractable true posterior $P_\theta(z|x)$ with some tractable parametric distribution $q_\phi(z|x)$. The marginal likelihood can be written as [15]

$$\log P_\theta(x) = D_{KL}[q_\phi(z|x)||P_\theta(z|x)] + \mathcal{L}(\theta, \phi; x) \tag{6}$$

where,

$$\mathcal{L}(\theta, \phi; x) = E_{q_\phi(z|x)}[\log P_\theta(x, z) - \log q_\phi(z|x)] \tag{7}$$

where, E and D_{KL} denotes the expectation and KL divergence respectively. The above marginal likelihood is again intractable due to KL divergence between approximate and true posterior, since we don't know true distribution. Since, $D_{KL} > 0$, $\mathcal{L}(\theta, \phi; x)$ is called as (variational) lower bound and act as a surrogate for optimizing the marginal likelihood. Re-parameterizing the random variable z and optimizing with respect to θ and ϕ yields [15],

$$\theta, \phi = \underset{\theta,\phi}{\text{argmax}} \quad \mathcal{L}(\theta, \phi : x) \approx \underset{\theta,\phi}{\text{argmax}} \quad \sum_{l=1}^{L} \log P_\theta(x|z^l) + D_{KL}[q_\phi(z^l|x)||P(z)] \tag{8}$$

Code and data: `github.com/anurendra/vae_sep`

where, θ and ϕ are the parameters of multi layered perceptrons (MLP) for encoders and decoders respectively, L denotes the total number of samples used in sampling. Often a single sample is enough for learning θ and ϕ, if we have enough training data [15]. Encoders and decoders are implemented via MLP networks with parameters θ and ϕ respectively. Normally, one layer neural network is used for encoders and decoder in VAE. However, number of layers can be increased for increasing the non-linearity. We call these as deep-VAE in the paper and show that deep-VAE performs better than VAE.

3. Source Separation

The audio single channel source separation (SCSS) aims to estimate the sources $s_i(t), \forall i$ from a mixed signal $y(t)$ made up of I sources, $y(t) = \sum_{i=1}^{I} s_i(t)$. We perform computations in the short time Fourier transform (STFT) domain. Given the STFT of the mixed signal $y(t)$, the primary goal is to estimate the STFT of each source $\hat{s}_i(t)$ in the mixture. Each of the sources is modeled using a single VAE i.e. a specific encoder and decoder for each source is learned. Fig. 2 shows the architecture of VAE used.

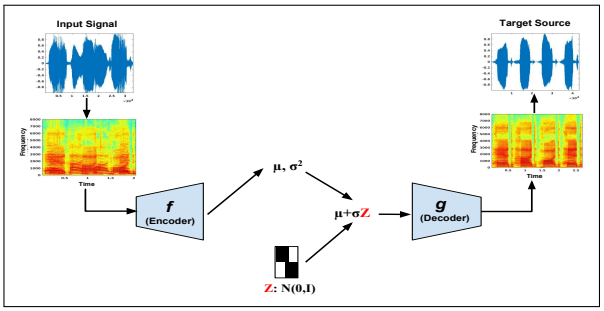


Figure 2: *Architecture of VAE for audio source separation*

We propose to use as many VAEs as the number of sources to be separated from the mixed signal. Each VAE deals with the mixed signal as a combination of a target source and background noise. These VAEs are trained to estimate the corresponding target sources from other background sources existing in the mixed signal. While training, VAEs map the magnitude spectrogram of the mixture to the magnitude spectrogram of the corresponding target sources. The inputs and outputs of the VAEs are 2D-segments from the magnitude spectrograms of the mixed and target signals respectively. This facilitates the VAEs capability to capture the time-frequency characteristics of each source by spanning multiple time frames.

3.0.1. *Training and Testing of VAEs for Source Separation*

Let's assume that we have training data as mixed signals and their corresponding clean sources. Let Y_{tr} be the magnitude spectrogram of the mixed signal and S_i be the magnitude spectrogram of the clean source i. The VAE of source i is trained to maximize the following likelihood function :

$$\theta, \phi = \underset{\theta,\phi}{\text{argmax}} \sum_{l=1}^{L} \log P_\theta(S_i|z^l) + D_{KL}[q_\phi(z^l|Y_{tr})||P(z)]$$

In practice, $L = 1$ in our set up leads to good learning of encoder and decoder. Given the trained VAEs, the magnitude spectrogram Y of the mixed signal is passed through all the

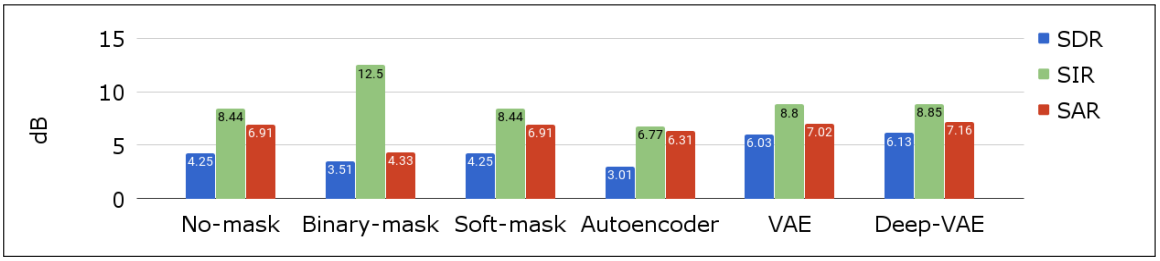


Figure 3: *Performance comparison of VAE ([513 128 64]) and deep VAE ([513 256 192 128 64]) with the baseline i.e. DNN and RNN with different masking function [7] and autoencoders [14].*

trained VAEs. The output of the VAE of source i is the estimate of the spectrogram $\tilde{S}_i$ of source i.

4. Experiments

In this section, we describe the experiments done for parameter selection of the model and it's comparative advantage over other existing deep learning architectures. We do parameter selection for latent dimension (K), batch size and encoder and decoder dimensions in VAE and deep VAE. We finally do the performance comparison of speaker source separation and show that VAE performs better or comparative to baseline methods.

4.1. Experimental Details

To validate the performance of the proposed model, we perform speaker source separation experiments on the TIMIT database [20]. To obtain the spectrogram from an audio signal, we perform short term Fourier transform (STFT) with 64 ms window and 16 ms overlap. Only magnitude of spectrogram were given as input to the algorithm. Finally, the separated time-domain speech was obtained by multiplying phase of the mixed signal with the magnitude of the separated spectrogram [2]. We have used a total of ten speakers (5 male and 5 female) from the database which includes speech data for five male and five female speakers sampled at 16 KHz. We normalize each of the signals to zero mean and unit variance. The training mixtures are obtained by linear addition of each male and female audio signals resulting in 25 mixtures at 0 dB signal to noise ratio (SNR). We trained five VAEs for five male and five female speakers respectively. The first 20 mixtures (4 for each male and female speaker) were used as input to VAE to train all networks for separation, and the last 5 mixtures were used for testing. For the input and output data for the VAEs, only magnitude of spectrogram was given as input to the encoder. We chose 17 frames as the number of spectral frames in each 2D-segment. So, the dimension of each input and output(target) for each VAE is 17 (time frames) * 513 (frequency bins). Finally, the separated time-domain speech was obtained by multiplying phase of the mixed signal with the magnitude of the separated spectrogram.

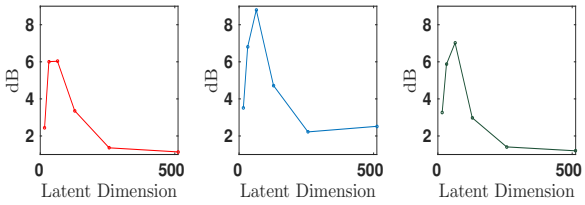
We use perceptually motivated scores as measure of evaluation and subsequently use BSS-EVAL TOOLBOX [21]. It proposes three metrics namely i) Source to distortion ratio (SDR) ii) Source to interference ratio (SIR) and iii) Source to artifact ratio (SAR).

4.2. Parameter Selection

The parameters of VAE were selected based on source separation results as evaluated on perceptual metrics described above.

4.2.1. Number of latent dimension(K)

We fixed all the dimensions except that of latent dimension K which was varied in [16 32 64 128 256 512]. The middle layer in encoder and decoder was fixed as 128, which was separately tuned.



(a) *SDR* (b) *SIR* (c) *SAR*

Figure 4: *Performance comparison for different latent space dimension [*16 32 64 128 256 512*]*

Fig. 4 shows the plot of source separation for a speaker in terms of SDR, SIR and SAR for different values of K. We see that model performs best for all the metrics when $K = 64$. Therefore we fix the latent dimension as 64 in later experiments.

4.2.2. Batch size

The magnitude spectrogram has temporal dependency along the direction of time. VAE learns the existing spatio-temporal structure from each batch of the spectrogram. There is a trade-off between increasing and decreasing batch size. As batch-size increases, VAE is able to extract long-term temporal features efficiently. However, it leads to loss of time specific structures. Also, lesser training data leads to a worse estimate of encoder and decoder. Very small batch size, on the other hand, does not allow VAE to learn the long-term temporal dependencies. Table 1 shows the performance of source separation as batch size is varied. Based on the results in the table, we fix batch size to be 17 in the rest of the experiments.

Table 1: *Performance of proposed framework for different batch size in terms of SDR, SIR and SAR.*

Batch size	Evaluation Metrics		
	SDR	SIR	SAR
1	1.21	1.66	1.16
10	2.46	3.31	2.66
17	**6.63**	8.80	7.02
30	6.61	8.92	6.95
50	6.73	**9.02**	6.92
70	5.61	8.07	**7.11**

4.2.3. Number of layers

The success of VAE lies in it's ability to learn the non-linear combination, and yet able to learn the posterior distributions efficiently. Therefore, we hypothesize that increasing number of layers in deep-VAE should yield better source separation. We use rectified linear units (ReLU) as the non-linearity everywhere in our network. We vary the number of layers keeping the latent dimension fixed to 64, found earlier. Table 2 shows the results. We see that the performance increases as the number of layer increases. However, deep architectures doesn't yield a substantial improvement given that these come at the cost of being more computationally expensive. The preference of Deep-VAE over VAE would, therefore, be dependent on the trade-off between accuracy and computational availability.

Table 2: *Performance of deep-VAE in terms of SDR, SIR and SAR.*

# Encoder Layers	Evaluation Metrics		
	SDR	SIR	SAR
[513 64]	2.04	3.18	2.01
[513 128 64]	6.03	8.80	7.02
[513 256 128 64]	6.13	**8.85**	7.16
[513 256 192 128 64]	**6.18**	8.84	**7.18**

4.3. Confidence Score

Unlike other existing source separation methods, VAE also yields posterior variance estimates of the separated sources. We calculate the average posterior variance as a proxy for the confidence scores of how good or bad are the separated sources. A lower variance implies that the distribution is peaky at mean and the confidence score is high. As discussed earlier, the signal to noise ratio for training data is (0 dB). For test signals, we vary the signal to noise ratio (SNR) and compute the average posterior variance. Fig. 5 shows the average posterior variance as SNR varies. It can be observed that average variance decreases as SNR increases (except at 0 dB). The anomaly at 0 dB can be attributed to the fact that VAE was trained on 0 dB SNR. Fig. 6 shows the reconstructed spectrogram in the case of different signal to noise ratios. We observe that the reconstruction spectrogram becomes noisy as SNR decreases.

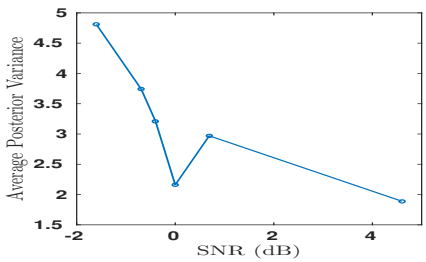


Figure 5: *Variation of average posterior variance with SNR*

4.4. Performance Analysis

We do perform analysis of VAE for source separation by comparing our algorithm with baseline approaches. We compare the performance of source separation with the existing deep learning approaches [7, 8] and autoencoders using a number of evaluation metrics. Table 3 shows the performance comparison of source separation of male and female individually with autoencoders. We see that VAE and deep-VAE performs better on

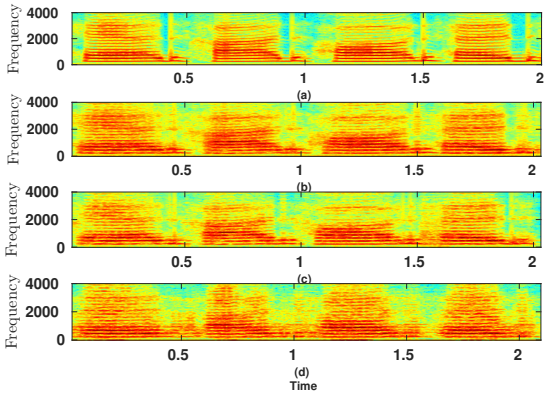


Figure 6: *(a): Target source (b): Reconstructed signal (SNR= 4.7 dB) (c): Reconstructed signal (SNR= 0 dB) (d): Reconstructed signal (SNR= -0.69 dB)*

all evaluation metrics with $\sim$ 3 dB improvement in SDR, $\sim$ 2 dB in SIR and $\sim$ 0.5 dB in SAR. Fig. 3 shows performance comparison of VAE and deep-VAE with baseline approaches. We see that VAE and deep-VAE perform best in terms of SDR with $\sim$ 2 dB improvement. In terms of SIR, Binary-mask approach performs the best. However, both VAE and deep-VAE provide good results in terms of all three measures. Note that for source separation, SDR would be considered the more important evaluation measure in which Binary-mask method performs far lower. While SDR captures overall noise, SIR and SAR captures only interference and artifact noise respectively [21]. This implies that VAE is able to remove noise (measured by SDR) better than all other models by capturing the spatio-temporal characteristic of spectrogram in latent space effectively. We also observe that SAR in VAE and deep-VAE is better than existing approaches. This shows that artifact introduced by VAE and deep VAE (measured by SAR) is lesser than other models.

Table 3: *Performance of proposed method in terms of SDR, SIR and SAR.*

Methods	Speakers	Evaluation Metrics		
		SDR	SIR	SAR
Autoencoders	Male	3.68	7.43	6.11
	Female	2.34	6.11	6.52
VAE	Male	6.26	8.91	7.27
	Female	5.93	8.74	6.77
Deep VAE	Male	**6.31**	**8.96**	**7.38**
	Female	**6.06**	**8.76**	**6.98**

5. Conclusions

In this work, we proposed a variational autoencoder based framework for monaural audio source separation. We showed that VAE is able to learn the inherent latent representation of a source by encoding the non-linear dependencies. The performance of the proposed framework is evaluated on audio source separation. The proposed framework yields reasonable improvements when compared to baseline methods. However, the framework requires prior knowledge of the sources in the mixture and a corresponding VAE has to be used (which allowed discriminative ability). Future works will be directed towards developing a single VAE for many/similar sources.

6. References

[1] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[2] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," *NIPS*, vol. 148, pp. 8–1, 2006.

[3] B. Raj, M. V. Shashanka, and P. Smaragdia, "Latent dirichlet decomposition for single channel speaker separation," in *ICASSP*, vol. 5. IEEE, 2006.

[4] N. Mohammadiha, P. Smaragdis, G. Panahandeh, and S. Doclo, "A state-space approach to dynamic nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 949–959, 2015.

[5] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[6] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative nmf and its application to single-channel source separation." in *INTERSPEECH*, 2014, pp. 865–869.

[7] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1562–1566.

[8] ——, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2136–2147, 2015.

[9] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[10] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[11] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[12] M. Kim and P. Smaragdis, "Adaptive denoising autoencoders: A fine-tuning scheme to learn from test mixtures," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 100–107.

[13] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[14] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," *arXiv preprint arXiv:1703.08019*, 2017.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[16] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.

[17] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," *arXiv preprint arXiv:1502.04623*, 2015.

[18] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[19] U. Jain, Z. Zhang, and A. Schwing, "Creativity: Generating diverse questions using variational autoencoders," *arXiv preprint arXiv:1704.03493*, 2017.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[21] C. Févotte, R. Gribonval, and E. Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.