City Research Online

City, University of London Institutional Repository

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/24240/

Link to published version: http://dx.doi.org/10.1109/icassp40776.2020.9053424

City Research Online: http://openaccess.city.ac.uk/   publications@city.ac.uk

IMPROVING SINGING VOICE SEPARATION WITH THE WAVE-U-NET USING MINIMUM HYPERSPHERICAL ENERGY

Joaquin Perez-Lapillo, Oleksandr Galkin, Tillman Weyde

Department of Computer Science
City, University of London
{joaquin.perez-lapillo,oleksandr.galkin,t.e.weyde}@city.ac.uk

ABSTRACT

In recent years, deep learning has surpassed traditional approaches to the problem of singing voice separation. The Wave-U-Net is a recent deep network architecture that operates directly on the time domain. The standard Wave-U-Net is trained with data augmentation and early stopping to prevent overfitting. Minimum hyperspherical energy (MHE) regularization has recently proven to increase generalization in image classification problems by encouraging a diversified filter configuration. In this work, we apply MHE regularization to the 1D filters of the Wave-U-Net. We evaluated this approach for separating the vocal part from mixed music audio recordings on the MUSDB18 dataset. We found that adding MHE regularization to the loss function consistently improves singing voice separation, as measured in the Signal to Distortion Ratio on test recordings, leading to the current best time-domain system for singing voice extraction.

Index Terms— Minimum Hyperspherical Energy, Music Source Separation, Deep Learning, Regularization

1. INTRODUCTION

Audio source separation is a core research area in audio signal processing with applications in speech, music and environmental audio. The main goal is to extract one or more target sources while suppressing other sources and noise [1]. One case is the separation of a singing voice, commonly presenting the main melody, from its musical accompaniment. Separating vocals from accompaniment has several applications, such as editing and remixing music, automatic transcription, generating karaoke tracks and music information retrieval [2].

We address this problem by applying Minimum Hyperspherical Energy (MHE) regularization to the Wave-U-Net model. In particular, our contributions are:

- A novel application of MHE to 1D convolutional neural networks with open source implementation.

- An extensive evaluation of MHE configurations and hyperparameters for singing voice separation.

- A new state-of-the-art for time-domain singing voice extraction on MUSDB18.

2. RELATED WORK

Approaches to source separation can be broadly divided into two groups: traditional and deep neural networks [3]. The first group contains techniques such as non-negative matrix factorization, Bayesian methods, and the analysis of repeating structures [4, 5, 6]. However, several deep neural network architectures have surpassed those models and achieve state-of-the-art performance today. [7] introduced the U-Net architecture for singing voice extraction, using spectrograms. The Wave-U-Net, proposed by [8], processes the audio data in the time domain, and is trained as a deep *end-to-end* model with state-of-the-art performance.

Many strategies for regularization have been developed, with early stopping, data augmentation, weight decay and dropout being the most commonly used [9]. Neural network based singing voice separation systems have used some of these methods in the past. [7] uses dropout. In [8], data augmentation is applied by varying the balance of the sources. Both methods also use early stopping.

In deep learning in general, overfitting has also been addressed by compressing the network [10], modifying the network architecture [11], and alleviating redundancy through diversification [12]. This last approach enforces diversity between neurons via regularization. Minimum Hyperspherical Energy, proposed by [13], also falls into this category. Inspired by a problem in physics [14], the MHE term aims to encourage variety between filters. MHE has increased performance in image classification, but the method has not yet been applied to audio.

3. METHOD

3.1. Architecture

Our work is based on the best Wave-U-Net configuration found by [8] for singing voice separation, which is a deep U-Net composed of 24 hidden convolutional layers, 12 in

each path, considering a filter size of 15 in the downsampling path and 5 for the upsampling. The first layer computes 24 feature maps, and each subsequent layer doubles that number. Each layer in the upsampling path is also connected to their corresponding layer in the downsampling path with skip-connections. The output layer generates an estimated waveform in the range [-1,1] for the vocals, while the accompaniment is obtained as the difference between the original mixture and the estimated singing voice, as in [8]. The original loss function is mean squared error (MSE) per sample.

3.2. Adding MHE to Wave-U-Net training

The MHE regularised loss function for the Wave-U-Net can be defined as:

$$Loss = MSE + \lambda_h * \sum_{j=1}^{L} \frac{1}{N_j(N_j - 1)} E_{sj}, \qquad (1)$$

where L is the number of hidden layers and N_j is the number of filters in the j-th layer. E_{sj} represents the hyperspherical energy of the j-th layer, given the parameter s, and it is calculated as:

$$E_{sj} = \sum_{i=1}^{N_j} \sum_{k=1, k \neq i}^{N_j} f_s(||\hat{w}_i - \hat{w}_k||), \qquad (2)$$

where $|| * ||$ is the Euclidean norm and $\hat{w}_i = \frac{w_i}{||w_i||}$ is the weight vector of neurons i projected onto a unit hypersphere. The dimensionality of the hypersphere is given by the number of incoming weights per neuron. For f_s, we use $f_s(z) = z^{-s}$ for $s > 0$, and $f_s(z) = \log z^{-1}$ otherwise [13].

With regard to λ_h, i.e. the weighting constant of the regulariser, [13] recommends to use a constant depending on the number of hidden layers, with the aim of reducing the weighting of MHE for very deep architectures. We use here $\lambda_h = \frac{1}{L}$ as defaul.

There are two possible configurations of MHE, full or half space. The half-space variation creates a *virtual* neuron for every existing neuron with inverted signs of the weights. The second term of equation 1 is then applied to $2N_j$ neurons in each hidden layer j. There are two alternative distance functions, Euclidean and angular. When dealing with angular distances, $||\hat{w}_i - \hat{w}_j||$ is replaced with $arccos(\hat{w}_i^T \hat{w}_j)$. The parameter s controls the scaling of MHE. Using the convention of [13] we have a total of twelve possible configurations as shown in Table 1.

Table 1. MHE configurations. Values of s with prefix a use angular distance. Each model can be used with each s [13].

Parameter	Values
model	[MHE, half_MHE]
s value	[0, 1, 2, a0 ,a1, a2]

3.3. Dataset

We use the MUSDB18 dataset [15][1]. It is composed of 150 full-length musical tracks. The data is divided into 100 songs for development and 50 for testing, corresponding to 6.5 and 3.5 hours, respectively. The development set is further divided into training and validation sets by randomly selecting 25 tracks for the latter.

For the task of singing voice separation, the drums, bass, and other sources are mixed to represent the accompaniment.

For the final experiment, we added the CCMixter dataset [16] featuring 50 more songs for training, as in [8]. In this setup, the last model is trained with a total of 150 songs. The test set remains unchanged.

3.4. Experimental setup

We explored the influence of the MHE configuration and hyperparameters with an initial grid search. For the grid search, all twelve MHE configurations described in Table 1 and a baseline model were implemented using the same network parametrization in order to compare their performance. Our implementation is available as open source[2].

As in [8], the models were trained using the Adam optimizer [17] with decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a batch size of 16 and learning rate of 0.0001. Data augmentation is applied by attenuating the vocals by a random factor in the range [0.7,1.0]. The best model selected in the first round of training is further fine-tuned, doubling the batch size and reducing the learning rate to 0.00001. The model with the lowest validation loss is finally tested against the unseen data.

In order to reduce the computational cost for the exploratory experiments, an epoch was defined as 1,000 iterations, instead of the original 2,000, and the early stopping patience (epochs without improvement of validation loss) was reduced from 20 to 10. Additionally, the tracks were mixed down to mono from stereo.

A further experiment was performed exploring the regularization constant $\lambda_h = 1/L$ and an increased early stopping patience. We finally evaluated the singing voice separation performance of the best MHE configuration with the original settings using the extended dataset and the parametrisation as in [8].

3.5. Evaluation

The models are evaluated using the signal-to-distortion ratio (SDR), as proposed by [18]. The audio tracks are partitioned into several non-overlapping segments of length one second to calculate SDR for each individual segment. The resulting SDRs are then averaged over the songs, and over the whole dataset. If For near-silent parts in one source the SDR can
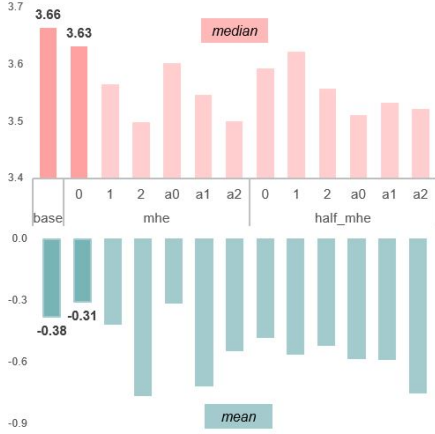
---

[1] https://sigsep.github.io/datasets/musdb.html
[2] https://github.com/jperezlapillo/Hyper-Wave-U-Net

Fig. 1. Median and mean test set SDR (in dB) for singing voice with different MHE loss configurations and patience 10.
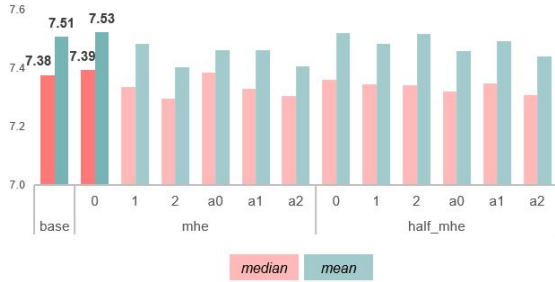

Fig. 2. Median and mean test set SDR (in dB) for accompaniment with different MHE losses and patience 10.

reach very low values, which can affect the global mean statistic. To deal with this issue, median statistics are also provided as in [8], along with standard deviation and median absolute deviation (MAD) for vocals and accompaniment.

4. RESULTS

4.1. Hyperparameter exploration

Figures 1 and 2 show statistics for the SDR obtained by each model over the MUSDB18 test set. The baseline model obtained the highest median SDR with 3.66 dB for the vocals estimation, followed closely by MHE_0 (full-space MHE, Euclidean distance, $s = 0$,) with 3.63 dB. However, MHE_0 obtained the highest mean SDR with -0.31 dB, compared to -0.38 dB for the baseline. For the accompaniment, the MHE_0 model showed the highest median and mean SDR, albeit by a small margin.

Table 2 compares different MHE configurations, aggregating the results of Figures 1 and 2 by groups of parameters. From this, it becomes clear that the ideal s value is 0 and the preferred distance is *Euclidean*. When comparing MHE
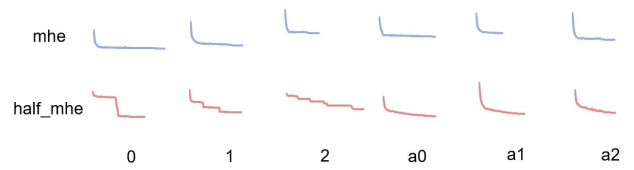

Fig. 3. MHE loss development during training (with patience 10). Longer curves represent longer training.

versus half_MHE models, the former obtains better results on vocals, and the latter on the accompaniment estimation.

4.2. MHE loss curves

Figure 3 shows the development of the MHE loss during training. While full MHE models rapidly reduce the MHE loss in the first epochs of the training process, Euclidean half_MHE models tend to form steps. The behaviour seems to increase in frequency when higher s values are in use.

The MHE loss tends to be relatively stable, with changes made every epoch in the range of $[0.0001, 0.00001]$. This is probably due to the large number of parameters being considered and the value chosen for learning rate (0.0001).

The following experiments focus on comparing the baseline model with the best performing MHE model, MHE_0, in different scenarios.

4.3. Early stopping and regularization strength

Table 3 shows the SDR results for MHE_0 with 20 epochs early stopping patience and different values for λ_h. This doubling of the early stopping patience led to a 88% increase in the number of epochs for the MHE_0 model. In this setting the MHE_0 model obtained higher results in all statistics compared to the baseline model on the leftmost column.

The increased early stopping patience leads to longer training times for the MHE_0 model with $\lambda_h = 1/L$, as recommended in in [13]. Table 3 shows the SDR results for $\lambda_h = 1/(2L)$ and $\lambda_h = 1$, too. It is clear that both increasing and decreasing the regularization constant have a negative effect on SDR.

4.4. Extended training data

Considering the same conditions reported by [8] for their best vocals separator system, called M4, we re-implemented this configuration and trained it in parallel with an MHE_0 model. The results in Table 4 show that MHE_0 outperforms our M4 implementation (left row) and the originally reported M4 results. This shows that MHE_0 is currently the best time-domain singing voice separation model.

Table 2. Average test set SDR results (in dB) aggregated by MHE hyperparameters.

		Model		s value			Distance	
		mhe	half_mhe	0	1	2	euclidean	angular
Vocals	Med	3.56	3.56	**3.58**	3.57	3.52	**3.58**	3.54
	MAD	2.85	2.90	2.89	2.87	2.86	2.87	2.87
	Mean	**-0.51**	-0.58	**-0.42**	-0.57	-0.64	**-0.51**	-0.58
	SD	13.81	14.02	13.78	13.99	13.97	13.89	13.94
Accomp.	Med	7.34	7.34	**7.37**	7.34	7.31	**7.35**	7.33
	MAD	2.11	2.11	2.11	2.11	2.10	2.10	2.11
	Mean	7.46	**7.49**	**7.49**	7.48	7.44	**7.49**	7.45
	SD	3.88	3.81	3.87	3.82	3.84	3.83	3.87

Table 3. Test set SDR (dB) with 20 epochs patience.

Value of λ_h		Basel.	1/(2L)	1/L	1
Voc.	Med	3.65	3.50	**3.69**	3.64
	MAD	3.04	2.82	2.98	2.96
	Mean	-0.56	-0.39	**0.01**	-0.10
	SD	14.23	13.53	13.31	13.30
Acc.	Med	7.37	7.32	**7.44**	7.42
	MAD	2.10	2.11	2.14	2.14
	Mean	7.53	7.44	**7.56**	7.54
	SD	3.74	3.92	3.89	3.89
Epochs		120	76	167	120

Table 4. Test set SDR (in dB) for the extended training set.

	Model	M4(reimp.)	MHE_0	M4(orig.)
Voc.	Med	4.44	**4.69**	4.46
	MAD	3.15	3.24	3.21
	Mean	0.17	**0.75**	0.65
	SD	14.38	13.91	13.67
Acc.	Med	10.54	**10.88**	10.69
	MAD	3.01	3.13	3.15
	Mean	11.71	**12.10**	11.85
	SD	6.48	6.77	7.03
Epochs		90	122	-

Fig. 4. Spectrograms of vocals with silent period.

5. DISCUSSION

The results confirm that the MHE_0 model, i.e. full-space MHE with Euclidean distance and s value= 0, outperforms the alternative MHE versions, similar to the results in [13]. The training for MHE models needs more epochs compared to the baseline and benefits from an increased early stopping patience. Overall, including MHE regularization in the Wave-U-Net loss function improves singing voice separation, when sufficient time is given for the training.

Particularly in the vocals source, there are some silent periods. These can produce very low SDR results and with very audible artifacts. The MHE helps to reduce artifacts in these periods, as can be seen in Figure 4.

The results achieved here do not yet match the best singing voice extracting method based on time-frequency representations [19]. The performance gap of around 1.5 dB is a research challenge that is worthwhile, because of the time-domain systems' potential for low latency processing.

6. CONCLUSION

This study explores the use of a novel regularization method, minimum hyperspherical energy (MHE), for improving the task of singing voice separation in the Wave-U-Net. It is, to the authors' knowledge, the first time that this technique is applied to an audio-related problem.

Our results suggest that MHE regularization, combined with the appropriate early stopping patience, is worth including in the loss function of deep learning separator systems such as the Wave-U-Net, as it leads to a new state of the art in our experiments.

For future work we intend to address other applications, such as speech enhancement and separation, as well as other loss formulations. We are also interested in designing low latency systems based on this approach and aim to reduce the computational cost of the Wave-U-Net.
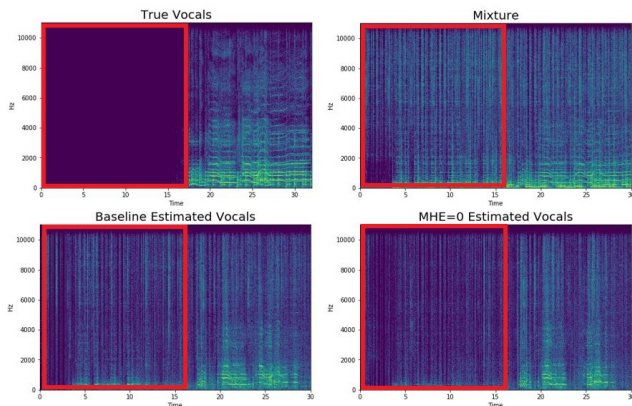
7. REFERENCES

[1] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, *Audio source separation and speech enhancement*, John Wiley & Sons, 2018.

[2] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel 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.

[3] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo, "An overview of lead and accompaniment separation in music," *arXiv:1804.08300 [cs, eess]*, 2018.

[4] Paris Smaragdis, Cedric Fevotte, Gautham J. Mysore, Nasser Mohammadiha, and Matthew Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[5] Alexey Ozerov, Pierrick Philippe, Frdric Bimbot, and Rmi Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1564–1578, 2007.

[6] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 73–84, 2013.

[7] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde, "Singing voice separation with deep u-net convolutional networks," *Paper presented at the 18th International Society for Music Information Retrieval Conference, 23-27 Oct 2017, Suzhou, China.*, 2017.

[8] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv:1806.03185 [cs, eess, stat]*, 2018.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016, http://www.deeplearningbook.org.

[10] Song Han, Huizi Mao, and William J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv:1510.00149 [cs]*, 2015.

[11] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861 [cs]*, 2017.

[12] Pengtao Xie, Jun Zhu, and Eric P. Xing, "Diversity-promoting bayesian learning of latent variable models," *arXiv:1711.08770 [cs, stat]*, 2017.

[13] Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song, "Learning towards minimum hyperspherical energy," *arXiv:1805.09298 [cs, stat]*, 2018.

[14] J.J. Thomson, "XXIV. on the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 7, no. 39, pp. 237–265, 1904.

[15] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "The MUSDB18 corpus for music separation," Dec. 2017, https://doi.org/10.5281/zenodo.1117372.

[16] Antoine Liutkus, Derry Fitzgerald, and Zafar Rafii, "Scalable audio separation with light kernel additive modelling," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 76–80.

[17] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[18] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji, "Open-unmix - a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, 2019.