Monaural Speech Separation with Deep Learning Using Phase Modelling and Capsule Networks

Toby Staines, Tillman Weyde and Oleksandr Galkin
Department of Computer Science
City, University of London
London, UK
{toby.staines, t.e.weyde, oleksandr.galkin}@city.ac.uk

***Abstract*—The removal of background noise from speech audio is a problem with high practical relevance. A variety of deep learning approaches have been applied to it in recent years, most of which operate on a magnitude spectrogram representation of a noisy recording to estimate the isolated speaking voice. This work investigates ways to include phase information, which is commonly discarded, firstly within a convolutional neural network (CNN) architecture, and secondly by applying capsule networks, to our knowledge the first time capsules have been used in source separation. We present a Circular Loss function, which takes into account the periodic nature of phase. Our results show that the inclusion of phase information leads to an improvement in the quality of speech separation. We also find that in our experiments convolutional neural networks outperform capsule networks at speech separation.**

***Index Terms*—Speech Separation, Speech Enhancement, Capsules, Phase, Convolutional Neural Networks**

I. Introduction

Source separation is the task of identifying and separating multiple sound sources from each other in a single mixed signal. Within this domain, speech separation (also referred to as speech enhancement or de-noising) focuses specifically on separating human voices from background noise, and has many modern-day applications, especially in the telecommunications and personal technology sectors. We focus on monaural audio, as in telephone signals, where one cannot take advantage of differing spatial origins of the sources.

The majority of existing work on this problem operates in the time-frequency domain to estimate the magnitude spectrogram of the isolated speech, which is then combined with the phase spectrogram of the original mixed signal. Estimation of the phase of an isolated source is difficult, not least because of the periodic nature of phase, but it is known to play an important role in source separation in the human auditory system [1] and has been demonstrated to aid with computational source separation in several pieces of recent work [2]–[5].

This work explores the estimation of a complete spectrogram in the time-frequency domain without computation with complex numbers, giving the advantage of being easily implementable with standard deep learning frameworks. Within an existing neural network architecture we directly estimate both magnitude and phase and/or the real and imaginary parts of the complex spectrogram. We also apply capsule networks; an approach which is motivated both by the parallel between the vector nature of a complex number and the vector output of a capsule, and by the previously demonstrated ability of capsule networks to identify overlapping characters in images [6].

The contributions of this work are:

1) A novel Circular Loss function for periodic variables.
2) The first application, to our knowledge, of capsule networks to auditory phase modelling.
3) Empirical evidence of improvement in speech separation by using phase information in various architectures.

II. Related Work

It was demonstrated by [2] that retaining phase information can lead to a reduction in artefacts (interference introduced by the separation process) in a musical source separation task, and [3] showed that combining the estimated magnitude with isolated rather than mixed signal phase results in better separation quality.

In another musical separation task, [4] include phase information from the mixed signal. They use the mixture phase to reconstruct the audio, giving positive results for some instrumental sources, but only a small improvement for voice, which is our target.

In [7], a deep neural network is trained to produce ideal ratio masks (IRMs) of the real and imaginary parts of the complex spectrogram, with some improvement over the magnitude estimation approach. [8] developed algorithmic components which work directly on the complex spectrogram and apply these in a Deep Complex Convolution LSTM network to speech separation on the TIMIT speech corpus, achieving a 2.3% improvement. However, their results are given in terms of mean squared error, which gives little indication of the audible quality of the results.

To address the periodicity of phase, [5] model the estimation as a classification problem by discretising phase values. This approach shows positive results when compared to a baseline model, which attempts to directly estimate phase as a continuous variable. Recently, [9] proposed the use of the instantaneous frequency for phase modelling, reaching an SDR value of 11.37 on the CHiME dataset.

We address periodicity with a novel loss function which supports a deep learning model to efficiently learn to estimate

periodic variables, such as phase, by regression. With this approach we investigate the performance of a number of different architectures to estimating the full complex spectrogram of the isolated speech signal.

We also explore the use of capsules, groups of neurons which produce a vector output, which have recently been developed for image processing tasks. In [6] they outperform a convolutional neural network (CNN) at identifying overlapping handwritten digits. This problem is analogous to separating the overlapping signals in a spectrogram, which motivates us to investigate the application of capsules to speech separation, the first time this has been attempted.

III. Method

A. Data and Evaluation Measures

The CHiME 3 dataset [10] consists of voice recordings made in a sound booth, and mixtures of these recordings with background noise from four different environments (a cafe, a bus, a pedestrian area and beside a busy road junction). The training set consists of 7138 utterances from 83 speakers, with an average length of 7.6 seconds, in total 15 hours of audio. The validation and test set contain 1620 and 1320 utterances by four other speakers each.

All models are evaluated using the standard source separation metrics defined in [11], using the implementation of [12]. Source to Distortion Ratio (SDR) measures all noise in the separated signal, Source to Artefact Ratio (SAR) measures the noise introduced by the separation and Source to Interference Ratio (SIR) measures noise from the original recording which has not been removed. We also report Normalised SDR (NSDR):

$$NSDR(s_m, s_r, s_e) = SDR(s_e, s_r) - SDR(s_m, s_r) \quad (1)$$

where s_m, s_r, and s_e are the mixed, reference and estimated sources, respectively. For all of these measures a higher score is better. The calculation of these metrics requires both source signals, but noise signals for specific mixes are not included in CHiME 3. We thus created mixes of the speech with random segments of the noise signals provided with CHiME 3.

B. Pre-Processing

The audio data has a sample rate of 16kHz and is converted to the time-frequency domain by short-time Fourier transform (STFT), with a window size of 1024 and a hop length of 256. Each utterance is then split into overlapping patches of 256 time frames (approximately four seconds) with a hop of 128, resulting in 50% overlap. The upper half of each spectrum is removed giving 512 frequency bins by 256 time periods. Magnitude is linearly re-scaled to a range of [0,1].

C. The Baseline Model

As a baseline we use the U-Net described in [13], a CNN model which showed state of the art performance in a musical source separation task, and is itself based on the original U-Net developed for medical image segmentation in [14]. It consists of a convolutional encoder and a transposed convolutional decoder, with corresponding layers in each connected by concatenation of their outputs. Layers in the encoder employ ReLu activation, while the decoder uses Leaky ReLu with 0.2 leakiness. Batch normalisation is employed throughout.

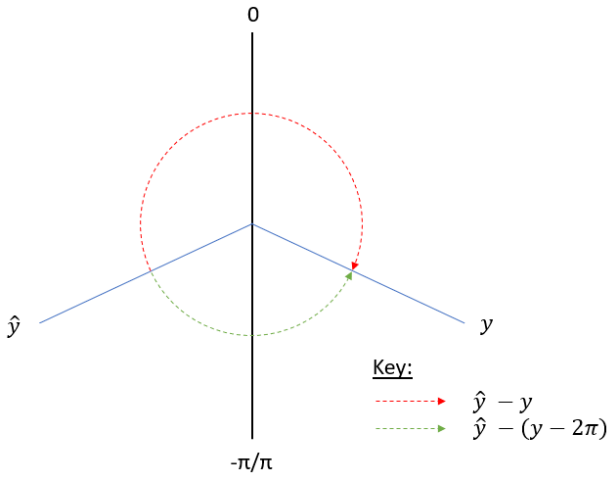

Fig. 1: The Circularity of Phase - The red line shows the direction of optimisation with a standard L_1 loss function, where $\hat{y}$ is the estimate of target value y. The green line shows the direction of optimisation with the Circular Loss function.

When presented with a mixed signal magnitude spectrogram X, the model estimates a ratio mask, which is multiplied element-wise with the input to provide the estimated magnitude of the isolated voice $\hat{Y}$:

$$\hat{Y} = f(X, \Theta) \odot X \quad (2)$$

where $f(X, \Theta)$ is the mask produced by the network with parameters Θ, when applied to X.

During training, the L_1 norm of the difference between the masked input and the target defines the loss. Formally, the magnitude loss, $L_m(X, Y; \Theta)$, of the network with parameters Θ is defined as:

$$L_m(X, Y; \Theta) = \|f(X, \Theta) \odot X - Y\|_1 \quad (3)$$

D. Circular Loss for Periodic Variables

Applying L_1 loss directly to a periodic variable such as phase is problematic, due to the discontinuity in the value which is being learned. Fig. 1 illustrates how a standard loss function, which assumes the error to be the target value minus the estimate, can lead to learning which adjusts the estimate in the direction away from the target on the unit circle. This is particularly problematic when the values are close to $+/-\pi$.

To address this issue we introduce the Circular Loss (L_c) which produces a suitable error signal by taking the error at each element of the estimated phase spectrogram as the lesser of the absolute value of the difference between the network output mask applied to the mixed input element x_{ij} and the target values y_{ij}, $y_{ij}+2\pi$ (i.e. forward one cycle), and $y_{ij}-2\pi$ (i.e. backward one cycle):

$$L_c(X, Y; \Theta) = \|P\|_1 \quad (4)$$

Where P is an $i \times j$ matrix where each element is defined by:

$$p_{ij} = \min(|f(x_{ij}, \Theta) \odot x_{ij} - y_{ij}|, \\ |f(x_{ij}, \Theta) \odot x_{ij} - (y_{ij} + 2\pi)|, \\ |f(x_{ij}, \Theta) \odot x_{ij} - (y_{ij} - 2\pi)|) \quad (5)$$

E. Data Representations

1) Magnitude: As a baseline, the network estimates a magnitude spectrogram mask, and the phase of the original mixture is used to estimate the isolated speech. We then attempt several alternative methods to incorporate the complex spectrogram information into the process. We evaluate different representations including redundant ones, to find which make the learning process more effective.

2) Phase Masking: Here magnitude and phase are encoded as a two channels (i.e. the input has shape [256, 512, 2]). The network produces an output of the same shape which is trained using L_m on the first channel and L_c on the second channel. The total loss is calculated as:

$$L = \frac{L_m(X, Y; \Theta) + W_c L_c(X, Y; \Theta)}{2}, \quad (6)$$

where the hyper-parameter W_c weights the Circular Loss.

The circular and magnitude losses are of very different magnitude; a W_c of 0.005 results in L_m and L_c being roughly equal at the outset of training.

3) Phase Difference: In this approach the data is presented in the same way, but the network is trained to estimate the phase difference between mixture and isolated voice; essentially an additive phase mask in the second channel. To accomplish this, a variation L_{pd} of Circular Loss is used:

$$L_{pd}(X, Y; \Theta) = \|f(X, \Theta) - D\|_1 \quad (7)$$

where D is an $i \times j$ matrix and

$$d_{ij} = \min(|x_{ij} - y_{ij}|, \\ |x_{ij} - (y_{ij} + 2\pi)|, \\ |x_{ij} - (y_{ij} - 2\pi)|) \quad (8)$$

4) Real and Imaginary: Here the two channels of the input data consist of the real and imaginary parts of the spectrogram and the network is trained to produce masks of the real and imaginary parts. For each channel L_1 loss is used and the real and imaginary parts are weighted equally.

5) Magnitude, Real and Imaginary: Here, we use a redundant representation with three channels for the network input and output. The real and imaginary output channels are used to calculate the phase. The loss is then calculated in the same way as for phase masking models.

6) Magnitude, Phase, Real and Imaginary: All four representations are provided to the network. The output has two channels, magnitude and phase, and the loss is calculated using equation 6.

7) Real and Imaginary to Magnitude and Phase: We use real and imaginary parts as in III-E4, but the output is a magnitude and phase mask as in subsection III-E2, so that the network models the relationship between real and imaginary values, and magnitude and phase.

TABLE I: Summary of Data Types - The input data representation and trained output for the data types described in III-E.

#	Input	Output
1	Magnitude	Magnitude mask
2	Magnitude & phase	Magnitude & phase masks
3	Magnitude & phase	Magnitude mask & phase difference
4	Real & imaginary	Real & imaginary masks
5	Magnitude, real & imaginary	Magnitude & phase masks
6	Magnitude, phase, real & imaginary	Magnitude & phase masks
7	Real & imaginary	Magnitude & phase masks

F. Capsule Models

We developed three capsule based architectures for speech separation; a simple proof of concept similar to the original capsule network in [6] and [15]'s baseline model, a CapsUNet, and a No-ConvCapsUNet. We use the Locally Constrained Dynamic Routing algorithm, developed for SegCaps in [15] for medical image segmentation to cope with the increased computation and memory costs caused by our data and network sizes compared to [6].

The Basic Capsule Network (BCN) consists of a single 128 filter convolutional layer, followed by two layers of capsules. For comparison against this model we also use a simple CNN, with three layers, each with same number of neurons as the BCN, but arranged and trained as a standard CNN.

The CapsUNet consists of a convolutional layer and four convolutional capsule layers in the encoder, with four deconvolutional capsule layers and three deconvolutional layers in the decoder, with skip connections (concatenations) between the layers on either side of the network.

The No-ConvCapsUNet is similar to the CapsUNet but the two channel input of real and imaginary parts is treated with a layer of capsules, rather than a convolutional layer.

G. Training Procedure

All models are trained using the ADAM optimiser [16] with an initial learning rate of 0.0001. U-Net models use a batch size of 50, but due to GPU memory constraints the simple capsule model uses a batch size of 10, and the CapsUNet and No-ConvCapsUNet use a batch size of five. An epoch is defined as one pass through the entire training set, and all models are trained for eight epochs. Training these models has significant computational cost, and stopping at this point allows experimentation with a wide range of parameters. Early experiments showed that validation set loss is stable by this point, and Circular Loss reaches a minimum after only two epochs.

For each data representation described above, apart from Magnitude and Real & Imaginary, U-Net models are trained using seven different W_c values, as shown in Table II. Each experiment was run three times and the results provided are the means of the three runs. Across all experiment settings there was a mean standard deviation in NSDR over the three runs of 1.1%.

TABLE II: Varying data representation with U-Nets. All results are means over three experiments. Green cells indicate a grater than 1% improvement relative to the baseline model. Red cells indicate a greater than 1% deterioration.

#	Data Type	Circular Loss Weighting (W_c)	Separation Metrics				Change Relative to Baseline (#1)			
			SDR	*SIR*	*SAR*	*NSDR*	*SDR*	*SIR*	*SAR*	*NSDR*
1	Magnitude	-	11.811	17.149	13.648	8.135	-	-	-	-
2	Phase Mask	0.5	11.757	15.978	14.165	8.081	-0.46%	-6.83%	3.79%	-0.66%
		0.05	11.861	16.539	13.994	8.185	0.42%	-3.56%	2.54%	0.61%
		0.005	11.961	17.150	13.836	8.285	1.27%	0.01%	1.38%	1.84%
		0.0005	**12.086**	17.000	14.073	**8.410**	**2.33%**	-0.87%	3.12%	**3.38%**
		0.00005	11.950	16.832	13.979	8.274	1.18%	-1.85%	2.43%	1.71%
		0.00001	12.047	16.902	14.064	8.371	2.00%	-1.44%	3.05%	2.90%
		0.000005	12.075	16.892	14.120	8.398	2.23%	-1.50%	3.46%	3.24%
3	Phase Difference	0.5	11.891	16.369	14.134	8.215	0.68%	-4.55%	3.56%	0.98%
		0.05	11.986	16.767	14.055	8.310	1.48%	-2.23%	2.99%	2.15%
		0.005	12.076	17.128	13.990	8.399	2.24%	-0.12%	2.51%	3.25%
		0.0005	11.980	16.721	14.076	8.304	1.43%	-2.49%	3.14%	2.08%
		0.00005	12.013	16.679	**14.144**	8.337	1.71%	-2.74%	**3.64%**	2.48%
		0.00001	11.999	16.932	13.992	8.323	1.59%	-1.27%	2.53%	2.31%
		0.000005	11.894	16.610	14.009	8.218	0.70%	-3.15%	2.65%	1.02%
4	Real & Imaginary	n/a	11.619	**19.074**	12.709	7.943	-1.63%	**11.22%**	-6.88%	-2.36%
5	Magnitude, Real & Imaginary	0.5	11.165	15.449	13.626	7.489	-5.47%	-9.91%	-0.16%	-7.94%
		0.05	11.927	17.076	13.808	8.251	0.98%	-0.43%	1.17%	1.42%
		0.005	11.959	17.188	13.790	8.283	1.25%	0.22%	1.04%	1.81%
		0.0005	11.905	17.051	13.781	8.229	0.79%	-0.58%	0.98%	1.15%
		0.00005	11.805	16.841	13.755	8.128	-0.06%	-1.80%	0.78%	-0.08%
		0.00001	11.912	16.801	13.916	8.262	0.85%	-2.03%	1.96%	1.56%
		0.000005	11.861	17.057	13.728	8.185	0.42%	-0.54%	0.59%	0.62%
6	Magnitude, Phase, Real & Imaginary	0.5	11.773	16.218	14.067	8.097	-0.33%	-5.43%	3.07%	-0.47%
		0.05	12.072	17.331	13.891	8.396	2.21%	1.06%	1.78%	3.21%
		0.005	11.978	17.164	13.855	8.302	1.41%	0.09%	1.52%	2.05%
		0.0005	12.021	17.008	13.967	8.345	1.77%	-0.82%	2.34%	2.58%
		0.00005	11.961	17.275	13.767	8.285	1.27%	0.73%	0.87%	1.84%
		0.00001	11.995	17.181	13.857	8.319	1.56%	0.18%	1.53%	2.26%
		0.000005	11.956	17.383	13.703	8.280	1.22%	1.36%	0.41%	1.78%
7	Real & Imaginary to Magnitude & Phase	0.5	11.834	16.471	13.985	8.158	0.19%	-3.95%	2.47%	0.28%
		0.05	12.024	17.247	13.856	8.348	1.80%	0.57%	1.52%	2.62%
		0.005	11.973	17.022	13.894	8.297	1.37%	-0.74%	1.80%	1.99%
		0.0005	11.895	16.788	13.911	8.219	0.71%	-2.11%	1.93%	1.02%
		0.00005	11.986	16.780	14.053	8.310	1.48%	-2.15%	2.97%	2.15%
		0.00001	12.020	16.979	13.982	8.344	1.77%	-1.00%	2.45%	2.57%
		0.000005	11.945	17.144	13.813	8.269	1.14%	-0.03%	1.21%	1.65%

IV. Results and Evaluation

A. Data Representation Experiments

Table II details the speech separation quality achieved by each of the data representations and Circular Loss weightings investigated. All five methods where W_c was varied showed a statistically significant improvement in NSDR over the baseline approach when $W_c \leqslant 0.05$, providing strong evidence that the inclusion of phase information does improve the quality of speech separation by deep learning. Significance was determined by a two sample t-test at the 5% confidence level between the three baseline experiments and 18 experiments for each data type. However, in most cases, overweighting phase with a W_c value of 0.5 led to a decrease in performance. Phase masking with a phase loss weight of 0.0005 provided the strongest results in terms of SDR and NSDR, with improvements of 2.33% and 3.38%, respectively. All Phase Mask and Difference models improved SAR, but that improvement came at a cost of worsened SIR in most cases, which changed by between 0.01% and -3.15%. However, according to [11], interferences tend to have less of an impact on perceived sound quality than artefacts and the overall effect as captured in SDR and NSDR is one of improvement in almost all models with phase. The converse of this is demonstrated by the Real & Imaginary model, where an 11.22% increase in SIR is observed against a SAR decrease of 6.88%. Although the NSDR only deteriorates by 2.36%, when listening to the output of these models the result is subjectively worse. The redundant inputs introduced in Magnitude, Phase, Real & Imaginary models led to slightly less improvement in SDR than Phase Mask and Difference models, but showed a better balance between SIR and SAR, in particular with $W_c = 0.05$, which is the only experiment with an improvement of over 1% in all metrics. Magnitude, Real & Imaginary and Real & Imaginary to Magnitude & Phase models both showed improvement in SAR and SDR, but not as strongly as other approaches.

TABLE III: Capsule Network Results

Model Architeture	Data Type	Separation Metrics			
		SDR	***SIR***	***SAR***	***NSDR***
U-Net	Magnitude	**11.811**	**17.149**	13.648	**8.135**
Caps U-Net	Magnitude	6.437	6.708	20.183	2.761
Caps U-Net	Phase Mask	6.279	6.680	18.398	2.603
BCN	Magnitude	7.317	10.924	10.674	3.641
Basic CNN	Magnitude	8.838	12.516	12.455	5.162
No-Conv Caps U-Net	Real & Imaginary	3.667	3.673	**38.408**	-0.009
No-Conv Caps U-Net	Real & Imaginary to Magnitude & Phase	4.561	4.876	19.695	0.885

B. Capsule Network Experiments

Both the BCN and CapsUNet architectures do show some success in speech separation and as far as we are aware this is the first time this has been demonstrated, but they do not perform at a level comparable to convolutional networks. The basic capsule network achieves a NSDR of 3.641, compared to 5.162 achieved by the basic CNN model. When the complexity of the model is increased in the CapsUNet results deteriorate further, with a NSDR of 2.761 failing even to match that of the simpler capsule model, reflecting the results of [17], which found limited improvement when adding complexity to the original capsule architecture. The No-ConvCapsUNet showed little ability to learn, with NSDRs close to zero indicating that the convolutional layer at the start of the network is a key element, and that capsules acting directly on the vector representation of the complex spectrogram is not effective.

Capsule networks take far longer to train than standard neural networks, and are also far slower at test time. When combined with their relatively poor performance this indicates that significant further work is required if they are to provide a practical alternative to, or improvement on, existing deep convolutional networks for source separation.

V. Conclusion

In this study we have evaluated a number of neural network architectures for the estimation of full complex spectrograms in speech separation and demonstrated that the estimation of phase leads to improved results in separating speech from background noise, when compared to the traditional approach of estimating only the isolated magnitude. A Circular Loss function for the estimation of periodic variables was introduced, which produces suitable error signals for periodic variables. We have also successfully demonstrated the first application of capsule networks to source separation, although the results do not match those of CNNs.

We propose further work in the area with subjective testing on human listeners to gain a more robust understanding of the perceived quality of the models' outputs, and an evaluation of Deep Complex Networks [8] that would enable a direct comparison to our results.

References

[1] K. V. Nourski and J. F. Brugge, "Representation of temporal sound features in the human auditory cortex," *Reviews in the Neurosciences*, vol. 22, no. 2, jan 2011. [Online]. Available: https://www.degruyter.com/view/j/revneuro.2011.22.issue-2/rns.2011.016/rns.2011.016.xml

[2] M. Dubey, G. Kenyon, N. Carlson, and A. Thresher, "Does Phase Matter For Monaural Source Separation?" in *Neural Information Processing Systems (NIPS)*, no. 31, 2017. [Online]. Available: http://arxiv.org/abs/1711.00913

[3] K. Paliwal, K. Wojcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[4] J. Muth, S. Uhlich, N. Perraudin, T. Kemp, F. Cardinaux, and Y. Mitsufuji, "Improving DNN-based Music Source Separation using Phase Features," *Computing Research Repository (CoRR)*, 2018. [Online]. Available: http://arxiv.org/abs/1807.02710

[5] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation," *Interspeech 2018*, no. September, pp. 2713–2717, 2018. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2018/abstracts/1773.html

[6] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules," in *Neural Information Processing Systems (NIPS)*, no. Nips, 2017. [Online]. Available: http://arxiv.org/abs/1710.09829

[7] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[8] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep Complex Networks," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: http://arxiv.org/abs/1705.09792

[9] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 27, no. 1, pp. 63–76, 2019.

[10] E. V. Jon Barker, Ricard Marxer and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 504–511.

[11] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[12] C. Raffel, B. Mcfee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "A Transparent Implementation of Common MIR Metrics," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[13] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing Voice Separation with Deep U-Net Convolutional Network," 2017.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCIA)*, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[15] R. LaLonde and U. Bagci, "Capsules for Object Segmentation," in *Conference on Medical Imaging with Deep Learning (MIDL)*, 2018. [Online]. Available: http://arxiv.org/abs/1804.04241

[16] D. P. Kingma and J. L. Ba, "Adam : A Method for Stochastic Optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[17] E. Xi, S. Bing, and Y. Jin, "Capsule Network Performance on Complex Data," *Computing Research Repository (CoRR)*, 2017. [Online]. Available: http://arxiv.org/abs/1712.03480