MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Hyperbolic Audio Source Separation

Petermann, Darius; Wichern, Gordon; Subramanian, Aswin Shanmugam; Le Roux, Jonathan

TR2023-017 April 14, 2023

Abstract

We introduce a framework for audio source separation using embeddings on a hyperbolic manifold that compactly represent the hierarchical relationship between sound sources and time-frequency features. Inspired by recent successes modeling hierarchical relationships in text and images with hyperbolic embeddings, our algorithm obtains a hyperbolic embedding for each time-frequency bin of a mixture signal and estimates masks using hyperbolic softmax layers. On a synthetic dataset containing mixtures of multiple people talking and musical instruments playing, our hyperbolic model performed comparably to a Euclidean baseline in terms of source to distortion ratio, with stronger performance at low embedding dimensions. Furthermore, we find that time-frequency regions containing multiple overlapping sources are embedded towards the center (i.e., the most uncertain region) of the hyperbolic space, and we can use this certainty estimate to efficiently trade-off between artifact introduction and interference reduction when isolating individual sounds.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2023

© 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

HYPERBOLIC AUDIO SOURCE SEPARATION

Darius Petermann[1,2], *Gordon Wichern*[1], *Aswin Subramanian*[1], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Indiana University, Department of Intelligent Systems Engineering, Bloomington, IN, USA

ABSTRACT

We introduce a framework for audio source separation using embeddings on a hyperbolic manifold that compactly represent the hierarchical relationship between sound sources and time-frequency features. Inspired by recent successes modeling hierarchical relationships in text and images with hyperbolic embeddings, our algorithm obtains a hyperbolic embedding for each time-frequency bin of a mixture signal and estimates masks using hyperbolic softmax layers. On a synthetic dataset containing mixtures of multiple people talking and musical instruments playing, our hyperbolic model performed comparably to a Euclidean baseline in terms of source to distortion ratio, with stronger performance at low embedding dimensions. Furthermore, we find that time-frequency regions containing multiple overlapping sources are embedded towards the center (i.e., the most uncertain region) of the hyperbolic space, and we can use this certainty estimate to efficiently trade-off between artifact introduction and interference reduction when isolating individual sounds.

Index Terms— audio source separation, hyperbolic space, speech, music, sound hierarchy

1. INTRODUCTION

A fundamental paradigm in deep learning-based audio source separation algorithms is the idea of applying a mask to a feature representation (e.g., a magnitude spectrogram or learned basis) of an audio mixture signal [1–4]. By inverting or decoding the masked feature representation, we obtain the isolated time-domain source signals. While techniques that learn feature encoders and decoders directly based on waveform signals have achieved impressive performance [3, 5], they lack interpretability compared to techniques based on time-frequency (T-F) representations such as the short-time Fourier transform (STFT) spectrogram [6, 7]. Among these, algorithms such as deep clustering [2] and deep attractor networks [8] learn an embedding vector for each T-F bin, and create masks using classifiers and/or clustering algorithms to assign embeddings to sources. A fundamental problem for these approaches then becomes how to best learn a discriminative embedding for each T-F bin.

In this work, we take inspiration from recent advances in modeling language [9–11], graphs [11–13], images [14–19]in hyperbolic space, and explore their relevance for audio source separation. Unlike Euclidean spaces, hyperbolic spaces have an inherent ability to infer hierarchical representations from data with very little distortion [20, 21]. Hierarchical and tree-like structures are ubiquitous in many types of audio processing problems such as musical instrument recognition [22, 23] and separation [24], speaker identification [25], music synthesis [26], and sound event detection [27, 28]. However, all of these approaches model the hierarchical information globally by computing a single embedding vector for an entire audio clip. Recent work in image segmentation [17] learns a hyperbolic embedding at the pixel level, and we take a similar approach by computing a hyperbolic embedding for each T-F bin of an audio mixture spectrogram, as illustrated in Fig. 1.

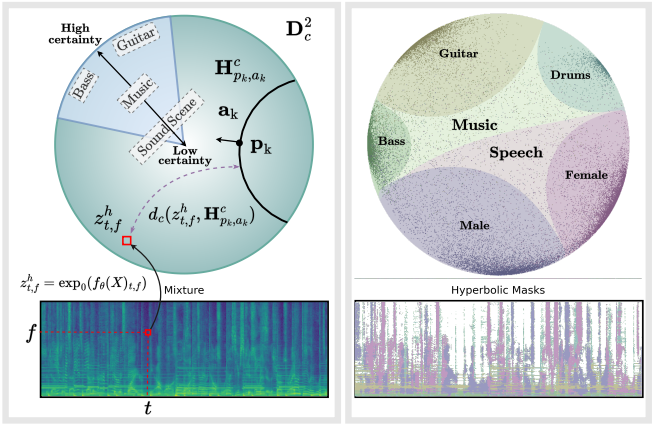

Fig. 1: Illustration of hyperbolic source separation. Left plot demonstrates the process of taking a T-F bin from a mixture spectrogram $X_{t,f}$ and projecting its Euclidean embedding to $z^h_{t,f}$ on the Poincaré ball (red squares); the dashed purple arrow represents the distance of the resulting embedding to the hyperplane $\mathrm{H}^c_{p_k,a_k}$. Right plot shows the Poincaré ball with the resulting decision boundaries from a trained model (top). The mixture embeddings (scatters) are colored-coded according to the maximum softmax layer output. The predicted T-F masks for each of the leaf sources (bottom) are plotted similarly, and silent bins (-40 dB and below) are colored in white.

As has been shown by many recent computer vision studies [14, 19], hyperbolic network layers [29] can be added to the end of existing network architectures to learn and classify hyperbolic embeddings. In addition to benefits in terms of hierarchical modeling, hyperbolic layers can learn strong embeddings at low dimensions [25, 26], represent uncertainty [14, 19], and estimate features such as edges in images based on their distance to the origin in the hyperbolic space [17]. Our main contribution is to explore whether similar observations are applicable to hyperbolic embeddings of T-F bins, and we propose new avenues for interpreting and interacting with hyperbolic embeddings in audio source separation applications.

Specifically, we use a simulated dataset containing multiple overlapping speakers and musical instruments to illustrate our approach as shown in Fig. 1. We classify each T-F bin embedding using hyperbolic softmax layers [29], and each T-F bin is further assigned to both a parent and child class using a hierarchical softmax [17]. We compare various widely used source separation loss functions, but find that energy-weighted cross-entropy performs best. Corroborating work in application domains such as computer vision and speaker identification, we find that performance with low embedding dimensions is relatively strong in hyperbolic space. In

This work was performed while D. Petermann was an intern at MERL.

terms of uncertainty, we show how the distance to the origin can be used to approximate computationally expensive certainty estimates obtained with Monte-Carlo dropout [30] using only a single forward pass. We also observe that T-F bins containing multiple overlapping sources are consistently embedded near the origin (i.e., they are considered the most uncertain embeddings). We then provide experiments demonstrating how this uncertainty can be exploited to navigate the trade-off between interference reduction and artifact introduction, which is fundamental to audio source separation.

2. HYPERBOLIC AUDIO SOURCE SEPARATION

2.1. Embedding-based source separation

We consider the problem of separating a mixture audio signal into K sources of interest. While we later focus more precisely on class-based separation, where the sources belong to K different sound classes, we first give a general presentation of separation approaches relying on time-frequency (T-F) embeddings.

Source separation is often formulated as a mask-inference problem [1] in a time-frequency domain, such as that obtained by a short-time Fourier transform (STFT), where one seeks to obtain the complex spectrogram $S^k \in \mathbb{C}^{T\times F}$ of each of the K sources by multiplying element-wise the spectrogram of the mixture $X \in \mathbb{C}^{T\times F}$ with a real-valued mask $M^k \in \mathbb{R}^{T\times F}$: $\hat{S}^k = M^k \odot X$. A common way to obtain the mask $M^k_{t,f}$ at time-frequency bin (t, f) relies on first computing an L-dimensional Euclidean embedding $z^{\mathrm{e}}_{t,f} \in \mathbb{R}^L$ from X using a deep neural network $f_\theta(.)$ parameterized by θ, such that $Z^{\mathrm{e}} = f_\theta(X) \in \mathbb{R}^{T\times F\times L}$. These embeddings can be projected to dimension K and passed through mask-output non-linearities such as sigmoid or softmax, or clustered in $\mathbb{R}^L$ to obtain a clustering of the T-F bins, such as in deep clustering [2]. In this work, we are interested in replacing these Euclidean embeddings and the way they are used to derive masks by equivalents in a hyperbolic space, as we hope that this leads to better embeddings that encompass hierarchical relationships between sounds. As a starting point, we focus here on the case of class-based separation with a softmax output layer, as this has been the primary non-linearity used in hyperbolic neural networks in computer vision [14, 17, 19], where the Euclidean embeddings and the multinomial logistic regression (MLR) classification approach applied to them are replaced by hyperbolic embeddings and a hyperbolic MLR. We leave the application to other non-linearities and to deep clustering to future work.

2.2. Hyperbolic embedding-based source separation

We first briefly review several notions pertaining to Riemannian manifolds and hyperbolic spaces that are useful in introducing hyperbolic embeddings and MLR, following [10, 29, 31].

A Riemannian manifold is defined as a pair consisting of a manifold $\mathcal{M}$ and a Riemannian metric g, where $g = (g_x)_{x\in\mathcal{M}}$ defines the local geometry g_x (i.e., an inner product) in the tangent space $T_x\mathcal{M}$ at each point $x \in \mathcal{M}$. While g defines the geometry locally on $\mathcal{M}$, it also defines the global shortest path, or *geodesic* (analogous to a straight line in Euclidean space), between two given points on $\mathcal{M}$. One can define an *exponential map* $\exp_x$ which projects any vector v of the tangent space $T_x\mathcal{M}$ onto $\mathcal{M}$, such that $\exp_x(v) \in \mathcal{M}$, and inversely a logarithmic map which projects any point in $\mathcal{M}$ back onto the tangent space at x.

The L-dimensional hyperbolic space is an L-dimensional Riemannian manifold of constant negative curvature $-c$. It can be described using several isometric models, among which we focus here on the Poincaré unit ball model $(\mathbb{D}^L_c, g^{\mathbb{D}}_c)$, defined in the space $\mathbb{D}^L_c = \{x \in \mathbb{R}^L \mid c||x||^2 < 1\}$. We assume $c > 0$, such that $\mathbb{D}^L_c$ corresponds to a ball of radius $\frac{1}{\sqrt{c}}$ in Euclidean space. Its Riemannian metric is given by $g^{\mathbb{D}}_c(x) = (\lambda^c_x)^2 g^E$, where $\lambda^c_x = 2/(1 - c\|x\|^2)$ is a so-called *conformal factor* and g^E the Euclidean metric. Given two points $x, y \in \mathbb{D}^L_c$, their induced hyperbolic distance d_c is obtained as

$$d_c(x, y) = \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\| - x \oplus_c y\|), \tag{1}$$

where $\oplus_c$ denotes the Möbius addition in $\mathbb{D}^L_c$, defined as

$$x \oplus_c y = \frac{(1 + 2c\langle x, y\rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c\langle x, y\rangle + c^2\|x\|^2\|y\|^2}. \tag{2}$$

One way to go back and forth between the Euclidean space $\mathbb{R}^L$ and the hyperbolic space $\mathbb{D}^L_c$ is to use the exponential and logarithmic maps at the origin 0, as $T_0\mathbb{D}^L_c = \mathbb{R}^L$, which can be obtained for $v \in \mathbb{R}^L \setminus \{0\}$ and $y \in \mathbb{D}^L_c \setminus \{0\}$ as:

$$\exp^c_0(v) = \frac{\tanh(\sqrt{c}\|v\|)}{\sqrt{c}\|v\|}v, \quad \log^c_0(y) = \frac{\tanh^{-1}(\sqrt{c}\|y\|)}{\sqrt{c}\|y\|}y. \tag{3}$$

We can thus obtain hyperbolic embeddings $z^h_{t,f} \in \mathbb{D}^L_c$ at the output of a classical neural network such as f_θ by simply projecting the usual Euclidean embeddings in that way:

$$z^h_{t,f} = \exp_0(z^e_{t,f}) = \exp_0(f_\theta(X)_{t,f}). \tag{4}$$

To define a hyperbolic softmax based on these hyperbolic embeddings, the Euclidean MLR can be generalized to the Poincaré ball as in [29]. In the Euclidean space, MLR is performed by considering the logits obtained by calculating the distance of an input's embedding $z \in \mathbb{R}^L$ (such as $z = z^e_{t,f}$) to each of K class hyperplanes, where the k-th class hyperplane is determined by a normal vector $a_k \in \mathbb{R}^L$ and a point $p_k \in \mathbb{R}^L$ on that hyperplane. Analogously, one can define a Poincaré hyperplane $\mathrm{H}^c_{p_k,a_k}$ by considering the union of all geodesics passing by a point p_k and orthogonal to a normal vector a_k in the tangent space $T_{p_k}\mathbb{D}^L_c$ at p_k. Hyperbolic MLR can then be defined by considering the distance from a hyperbolic embedding $z = z^d_{t,f} \in \mathbb{D}^L_c$ to each $\mathrm{H}^c_{p_k,a_k}$, leading to the following formulation as shown in [29]:

$$p(\kappa{=}k|z) \propto \exp\Big(\frac{\lambda^c_{p_k}\|a_k\|}{\sqrt{c}}\sinh^{-1}\Big(\frac{2\sqrt{c}|\langle -p_k \oplus_c z, a_k\rangle|}{(1-c\|-p_k\oplus_c z\|^2)\|a_k\|}\Big)\Big). \tag{5}$$

We can use the probability $p(\kappa = k|z^d_{t,f})$ that T-F bin (t, f) is dominated by the k-th source to obtain K source-specific mask values for each T-F bin in the input spectrogram X. This procedure is illustrated in Fig. 1. Note that p_k and a_k both parameterize the k-th hyperbolic hyperplane and are trainable. All parameters of the network f_θ and of the hyperplanes can be optimized using classical source separation objective functions, either on the masks or on the reconstructed signals.

2.3. Uncertainty in Audio Source Separation

It has been shown repeatedly [14, 17, 29, 31] that the distance $d_c(0, z^h_{t,f})$ of the projected embeddings to the center of the Poincaré ball can serve as a reliable measure of classification certainty. This hyperbolic distance can be computed from the L_2 norm of $z^h_{t,f}$ as considered in the Euclidean space using the monotonous relationship $\log((1 + c\|z^h_{t,f}\|^2)/(1 - c\|z^h_{t,f}\|^2))$, so that the L_2 norm can equivalently be considered as a measure of certainty. We here aim at validating this hypothesis with audio embeddings and bringing further light onto the notion of certainty in the context of audio source separation. While hyperbolic uncertainty has previously been exploited towards the network optimization stage [25], the direct impact of this notion on quality criteria such as amount of

artifacts, distortion, or interferences present in resulting signal has yet to be explored.

Furthermore, following the observation presented in [17], we argue that the notion of certainty observed on audio embeddings provides meaningful insights of signal content, such as the complexity of the scene or the number of active sources it may contain.

3. EXPERIMENTAL SETUP

LibriSpeech Slakh Unmix (LSX): To test our proposed method we built a simple hierarchical source separation dataset containing mixtures from two "parent" classes - *music* and *speech*, and five "leaf" classes - *bass*, *drums*, *guitar*, *speech-male*, and *speech-female*. As building blocks, we use the clean subset of LibriSpeech [32], and Slakh2100 [33], which is a dataset of 2,100 synthetic musical mixtures and containing bass, drums, and guitar stems in addition to various other instruments. We built a dataset consisting of 1947 mixtures, each 60 s in length, for a total of about 32 hours. The data splits are 70%, 20%, and 10% for training, validation, and testing sets, respectively. The *speech-male* and *speech-female* source targets consist of utterances randomly picked (without replacement) and concatenated consecutively (without overlap) until the 60 s track length is reached. Any signal from the last concatenated utterance exceeding that length was discarded. For Slakh2100, we only selected the first 60 s of the *bass*, *drums*, and *guitar* stems for each track. Any tracks with a duration less than 60 s were discarded. All sources were summed using their original gain to make the overall mixture. This led to challenging input SDR values ("No Proc." in Table 1), with standard deviation values ranging from 2-13 dB depending on the class.

Network architecture and training setup: Our model consists of four BLSTM layers with 600 units in each direction, followed by a dense layer to obtain an L-dimensional Euclidean embedding for each T-F bin. A dropout of 0.3 is applied on the output of each BLSTM layer, except the last. For the hyperbolic models ($c > 0$), an exponential projection layer is placed after the dense layer, mapping the Euclidean embeddings onto the Poincaré ball with curvature $-c$. As discussed in Section 2, MLR layers, either Euclidean or hyperbolic with softmax activation functions are then used to obtain masks for each of the source classes. In practice, we follow the hierarchical softmax approach from [17], and have two MLR layers: one with $K=2$ for the parent (speech/music) sources, and a second with $K=5$ for the leaf classes. We use the mixture phase for resynthesis and compare multiple training objectives in Section 4.

We use the ADAM optimizer for Euclidean parameters, and the Riemannian ADAM [34] implementation from *geoopt* [35] for hyperbolic parameters. All models are trained using chunks of 3.2 s and a batch size of 10 for 300 epochs using an initial learning rate of 10^{-3}, which is halved if the validation loss does not improve for 10 epochs. We use an STFT size of 32 ms with 50% overlap and square-root Hann window.

4. EXPERIMENTAL ANALYSIS

Model comparisons: Table 1 presents the scale-invariant signal-to-distortion ratio (SI-SDR) [36] on the LSDX dataset described in Sec. 3. We include the no processing condition (lower bound, using the mixture as estimate) and oracle phase sensitive mask [1] (upper bound). For loss functions, we use the phase-sensitive approximation (PSA) [1] with L_1 loss [37], and the waveform approximation (WA) loss training through the iSTFT [38]. Additionally, to compare

Table 1: SI-SDR in dB on the LSX test set with different loss functions and embedding dimensions L. The "Hyp." column denotes whether the model was trained using hyperbolic ($c = 1$) or Euclidean embeddings.

			Parents		Music Leaves			Speech Leaves		
Loss	L	Hyp.	Music	Speech	Bass	Drums	Guitar	Male	Female	Avg.
No Proc.			-2.9	3.0	-8.8	-11.8	-8.7	-3.2	-4.2	-5.2
$\text{Oracle}_{\text{PSF}}$			10.3	13.8	5.6	9.8	7.0	10.1	11.1	9.7
PSA	2	×	7.7	11.3	2.7	4.5	**3.0**	6.2	6.5	6.0
WA	2	×	7.1	10.8	**2.8**	3.2	2.3	5.5	5.6	5.3
CE_{IBM}	2	×	5.1	8.8	-2.3	1.8	-1.0	2.8	2.9	2.6
$\text{CE}_{\text{IBM, W.}}$	2	×	**8.0**	**11.5**	2.7	**4.9**	2.5	**6.4**	**6.6**	**6.1**
PSA	2	✓	7.5	10.9	-4.3	-6.9	-4.6	6.0	6.3	2.1
WA	2	✓	7.3	11.0	2.5	3.7	2.3	5.3	5.4	5.4
CE_{IBM}	2	✓	5.4	9.0	-0.3	2.6	0.8	3.4	3.8	3.5
$\text{CE}_{\text{IBM, W.}}$	2	✓	**7.9**	**11.4**	**2.9**	**5.1**	**3.2**	**6.5**	**7.0**	**6.3**
PSA	128	×	7.7	11.2	**3.3**	5.0	3.3	6.1	6.4	6.1
WA	128	×	7.4	11.2	**3.3**	4.8	3.0	5.9	6.1	6.0
CE_{IBM}	128	×	6.5	10.1	1.7	6.1	2.5	5.1	5.7	5.4
$\text{CE}_{\text{IBM, W.}}$	128	×	**7.9**	**11.5**	**3.3**	**6.8**	**3.8**	**6.8**	**7.3**	**6.8**
PSA	128	✓	7.5	10.9	2.9	4.9	3.1	5.8	6.0	5.9
WA	128	✓	7.0	11.2	**3.3**	4.8	2.9	5.9	6.0	5.9
CE_{IBM}	128	✓	6.0	9.7	1.3	5.8	2.0	4.5	5.2	4.9
$\text{CE}_{\text{IBM, W.}}$	128	✓	**7.8**	**11.3**	**3.3**	**6.1**	**3.7**	**6.7**	**7.1**	**6.6**

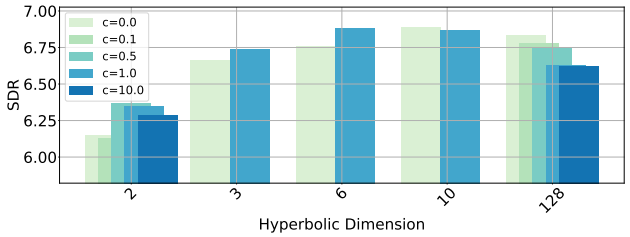

Fig. 2: SI-SDR [dB] averaged across all (parent and leaf) sources for various curvature values and embedding dimensions.

with related image segmentation approaches [17], we explore using a cross entropy (CE) loss with the ideal binary mask (IBM) as a training target [4], denoted by CE_{IBM} in Table 1, and a magnitude-ratio-weighted CE loss inspired by the deep clustering weighting function from [37], denoted by $\text{CE}_{\text{IBM, W.}}$, where the weight for each T-F bin is set to the ratio of the magnitude of the mixture at that bin to the sum over all bins.

In Table 1, CE_{IBM} consistently performs poorly, likely because low-energy and high energy T-F bins are weighted equally. However, $\text{CE}_{\text{IBM, W.}}$ consistently outperforms PSA and WA, which we hypothesize is because the CE objective is particularly well-matched to the softmax mask nonlinearity we use in this work. We also note that, consistent with previous hyperbolic work [17, 25, 26, 29, 39], performance of hyperbolic models is relatively stronger than Euclidean ones at low ($L = 2$) embedding dimension, as they allow for efficiently packing hierarchical structure, but that advantage fades away as embedding dimension increases. This result is further confirmed in Fig. 2, where we explore multiple embedding dimensions and curvature parameters. Unless otherwise stated, the hyperbolic configuration used in all subsequent experiments uses the $\text{CE}_{\text{IBM, W.}}$ loss, curvature parameter $c = 0.1$, and embedding dimension $L = 2$.

Hyperbolic Certainty and Audio Semantics: In Fig. 3, we show that embeddings associated with T-F bins containing many active sources (i.e., 4+) tend to be positioned close to the center of the Poincaré Ball, where the certainty is low. In contrast, embeddings with a single active source tend to be located closer to the edge of the Poincaré Ball, where the certainty is high. Intuitively, the higher the number of sound classes (i.e., interferences) active, the more uncertain as for what specific single source the bin embedding may

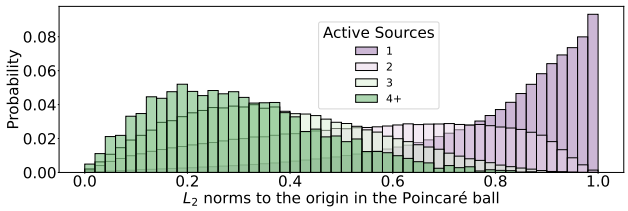

Fig. 3: Histograms denoting the distribution of L_2 norms to the origin of the Poincaré Ball for all embeddings in our test-set. We define the number of active sources for each embedding based on the ground-truth mask values of their associated T-F bin.

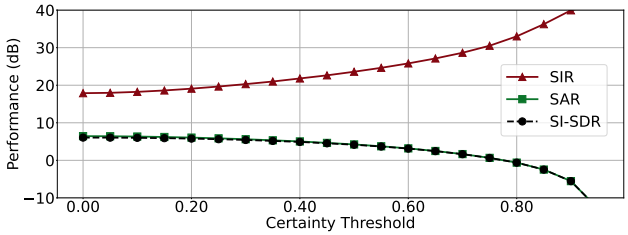

Fig. 4: Separation metrics as a function of certainty thresholding (here applied to the L_2 norm $\|z^h_{t,f}\|_2$). A threshold of 0 means all hyperbolic embeddings are taken into account during inference, and a threshold of 0.95 means only embeddings with high certainty (i.e., high distance from the origin) are considered, while bins below the threshold are silenced.

belong to. We consider a source active for a given T-F bin if it satisfies the following conditions: (1) it is non-silent, i.e., it is not more than 20.0 dB below the maximum T-F bin energy in a given sound file, and (2) the ratio of the ground-truth T-F bin magnitude for that source divided by the sum of ground-truth magnitudes for all sources (i.e., the ideal ratio mask) is greater than 0.1.

Hyperbolic Distance - Interference and Artifacts: We also explore a simple approach for using hyperbolic certainty estimates to modify and interact with mask values. Specifically, we embed all T-F bins for a given mixture in the Poincaré ball, and any bins that are close to the origin (i.e., low certainty) have their mask values set to zero. Figure 4 shows the impact of varying this certainty threshold on the SI-SDR, SIR, SAR [36, 40] metrics. Each data point denotes an evaluation pass on the entire test set given a certainty threshold (x-axis). We see that if we silence uncertain T-F bins in the mask predictions, the resulting signals will contain less interference at the expense of an increase in artifacts. Using hyperbolic certainty to control the trade-off between artifact introduction and interference reduction could be an exciting area to further explore in future work.

Comparison with Bayesian certainty: To further validate our hyperbolic certainty measurements, we lead an experiment comparing hyperbolic and Bayesian certainty. The latter approach is generally achieved by means of Monte-Carlo dropout [41, 42] where multiple stochastic forward passes using a trained network with different dropout realizations are performed to obtain the posterior distribution for a single input example. The network prediction and its associated certainty map can then be inferred from the mean and variance computed over all stochastic passes. In order to deal with the fact that our network makes predictions over multiple sound classes, we follow the approach presented in [30] and formulate the Bayesian certainty $\zeta_{t,f}$ at T-F bin (t, f) as the negative predictive entropy

$$\zeta_{t,f} = \sum_{k=1}^{K_{\text{leaf}}} \Big(\frac{1}{N} \sum_{n} p(\kappa = k | \hat{z}^{(n)}_{t,f}) \Big) \log \Big(\frac{1}{N} \sum_{n} p(\kappa = k | \hat{z}^{(n)}_{t,f}) \Big), \quad (6)$$

where K_{leaf} denote the number of leaf classes (i.e., without the parent

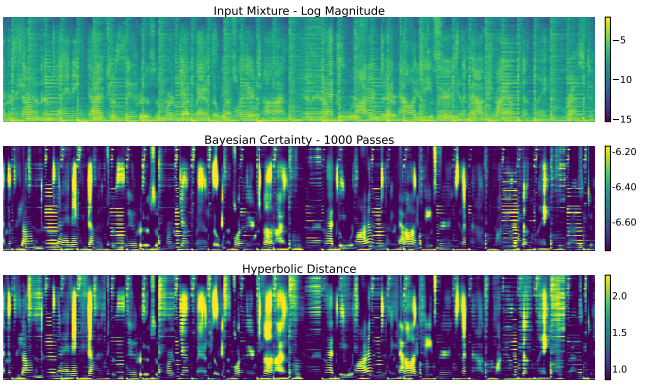

Fig. 5: Comparison between Bayesian and hyperbolic certainty. The Bayesian certainty is computed by running 1000 passes with Monte-Carlo dropout during inference and computing the bin-wise negative entropy across all passes on the five leaf nodes (i.e., excluding the parent nodes). The hyperbolic certainty is obtained by computing the hyperbolic distance from the Poincaré ball origin $d_c(0, z^h_{t,f})$ for each T-F bin embedding $z^h_{t,f}$.

classes), N the total number of stochastic passes, and $p(\kappa = k|\hat{z}^{(n)}_{t,f})$ the probability, expressed as the softmax output, that T-F bin (t, f) belongs to source class k, and $\hat{z}^{(n)}_{t,f}$ is the embedding obtained with the dropped out network parameters for stochastic pass n. Following [30], we set the dropout rate to 0.5 at the output of each BLSTM layer of our network. A low entropy (i.e., high certainty) means that the predictability is high, while a high entropy (low certainty) means that mask values fluctuate significantly across passes.

Fig. 5 provides a visual contrast of the certainty map computed using 1000 Monte-Carlo dropout iterations and the one obtained using the hyperbolic distance from the Poincaré ball origin for each T-F embedding in a single forward pass. The colorscale limits on both maps are set using the 30th and 95th percentiles to ensure a fair qualitative comparison, even though the scales are different. We observe a clear resemblance between both maps. The correlation coefficient between the two certainty maps is $\rho = 0.75$. This observation tells us that the hyperbolic certainty map, which comes for free (i.e., $N = 1$), can be as interpretable as its Bayesian counterpart, which requires multiple forward passes at inference time.

Code and demo are publicly available [1].

5. CONCLUSION

We have investigated the use of the Poincaré Ball model to perform audio source separation in the hyperbolic space. Our hyperbolic model operates and computes T-F embeddings in the Euclidean space and projects them onto the hyperbolic space. Masks are obtained by hyperbolic multinomial logistic regression considering the distance from hyperbolic embeddings to hyperbolic hyperplanes. Through our experimental setup, we have demonstrated that hyperbolic audio embeddings can convey useful information in regards to uncertainty and underlying hierarchical sound structures. We associated these notions to known audio concepts such as artifacts and interferences. In the future, we aim at exploring how the hyperbolic space, and especially the notion of hierarchy, could benefit additional audio-related tasks, such as audio tagging and sound-event detection. In the context of audio source separation, we believe that further connections to deep clustering, and more complex and deeper sound taxonomies are worth exploring.

[1] https://github.com/merlresearch/hyper-unmix/

6. REFERENCES

[1] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, Mar. 2016.

[3] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[4] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[5] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proc. ISMIR Workshop on Music Source Separation*, 2021.

[6] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude *et al.*, "Demystifying TasNet: A dissecting approach," in *Proc. ICASSP*, May 2020.

[7] R. Parikh, I. Kavalerov, C. Espy-Wilson, and S. Shamma, "Harmonicity plays a critical role in DNN based versus in biologically-inspired monaural speech segregation systems," in *Proc. ICASSP*, May 2022.

[8] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 787–796, 2018.

[9] A. Suzuki, A. Nitanda, J. Wang, L. Xu *et al.*, "Generalization error bound for hyperbolic ordinal embedding," in *Proc. ICML*, 2021.

[10] R. Shimizu, Y. Mukuta, and T. Harada, "Hyperbolic neural networks++," in *Proc. ICLR*, 2021.

[11] J. Sun, Z. Cheng, S. Zuberi, F. Perez *et al.*, "HGCF: Hyperbolic graph convolution networks for collaborative filtering," in *Proc. WWW*, 2021.

[12] Q. Liu, M. Nickel, and D. Kiela, "Hyperbolic graph neural networks," in *Proc. NeurIPS*, Dec. 2019.

[13] G. Bachmann, G. B'ecigneul, and O.-E. Ganea, "Constant curvature graph convolutional networks," in *Proc. ICML*, Jul. 2020.

[14] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets *et al.*, "Hyperbolic image embeddings," in *Proc. CVPR*, Jun. 2020.

[15] I. Chami, A. Gu, V. Chatziafratis, and C. Ré, "From trees to continuous embeddings and back: Hyperbolic hierarchical clustering," in *Proc. NeurIPS*, Dec. 2020.

[16] Y. Guo, X. Wang, Y. Chen, and S. X. Yu, "Free hyperbolic neural networks with limited radii," in *Proc. CVPR*, Jun. 2022.

[17] M. G. Atigh, J. Schoep, E. Acar, N. van Noord *et al.*, "Hyperbolic image segmentation," in *Proc. CVPR*, Jun. 2022.

[18] Z. Weng, M. G. Ogut, S. Limonchik, and S. Yeung, "Unsupervised discovery of the long-tail in instance segmentation using hierarchical self-supervision," in *Proc. CVPR*, 2021.

[19] D. Surís, R. Liu, and C. Vondrick, "Learning the predictability of the future," in *Proc. CVPR*, 2021.

[20] F. Sala, C. De Sa, A. Gu, and C. Re, "Representation tradeoffs for hyperbolic embeddings," in *Proc. ICML*, Jul. 2018.

[21] R. Sarkar, "Low distortion delaunay embedding of trees in hyperbolic plane," in *Proc. International Symposium On Graph Drawing*, 2012.

[22] S. Essid, G. Richard, and B. David, "Hierarchical classification of musical instruments on solo recordings," in *Proc. ICASSP*, May 2006.

[23] H. F. Garcia, A. Aguilar, E. Manilow, and B. Pardo, "Leveraging hierarchical structures for few-shot musical instrument recognition," in *Proc. ISMIR*, Nov. 2021.

[24] E. Manilow, G. Wichern, and J. Le Roux, "Hierarchical musical instrument separation." in *Proc. ISMIR*, Oct. 2020.

[25] J. Lee, K. Sung-Bin, S. Kang, and T.-H. Oh, "Lightweight speaker recognition in Poincaré spaces," *IEEE Signal Process. Lett.*, vol. 29, pp. 224–228, 2022.

[26] F. Nakashima, T. Nakamura, N. Takamune, S. Fukayama *et al.*, "Hyperbolic timbre embedding for musical instrument sound synthesis based on variational autoencoders," in *Proc. APSIPA ASC*, 2022.

[27] A. Jati, N. Kumar, R. Chen, and P. Georgiou, "Hierarchy-aware loss function on a tree structured label space for audio event detection," in *Proc. ICASSP*, May 2019.

[28] J. Cramer, V. Lostanlen, A. Farnsworth, J. Salamon *et al.*, "Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers," in *Proc. ICASSP*, May 2020.

[29] O. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," in *Proc. NeurIPS*, vol. 31, 2018.

[30] J. Mukhoti and Y. Gal, "Evaluating bayesian deep learning methods for semantic segmentation," *arXiv preprint arXiv:1811.12709*, 2018.

[31] W. Peng, T. Varanka, A. Mostafa, H. Shi *et al.*, "Hyperbolic deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 12 2021.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, Apr. 2015.

[33] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proc. WASPAA*, Oct. 2019.

[34] H. Kasai, P. Jawanpuria, and B. Mishra, "Riemannian adaptive stochastic gradient algorithms on matrix manifolds," in *Proc. ICML*, Jun. 2019.

[35] M. Kochurov, R. Karimov, and S. Kozlukov, "Geoopt: Riemannian optimization in PyTorch," *arXiv preprint arXiv:2005.02819*, 2020.

[36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. ICASSP*, May 2019.

[37] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. ICASSP*, Apr. 2018.

[38] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, Sep. 2018.

[39] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Proc. NeurIPS*, vol. 30, 2017.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1462 – 1469, Aug. 2006.

[41] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, Jun. 2015.

[42] P. Nazreen and A. G. Ramakrishnan, "DNN based speech enhancement for unseen noises using Monte Carlo dropout," in *Proc. ICSPCS*, 2018, pp. 1–6.