

FIA 2020/22

XII CONGRESSO/CONGRESO IBEROAMERICANO DE ACÚSTICA
XXIX ENCONTRO DA SOCIEDADE BRASILEIRA DE ACÚSTICA - SOBRAC

Florianópolis, SC, Brasil

A deep learning approach for high resolution BRIRs generation for real-time AVR.

Sanaguano-Moreno, D.A.¹; Lucio-Naranjo, J.F.²; Tenenbaum, R.A.³; Regattiere-Sampaio, G.B.⁴

¹⁻² Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito, Pichincha, Ecuador,
{daniel.sanaguano,jose.lucio}@epn.edu.ec

³ Pós-Graduação em Engenharia Civil, Universidade Federal de Santa Maria, Santa Maria, RS, Brasil, ratenenbaum@gmail.com

⁴ Engenharia Acústica, Universidade Federal de Santa Maria, Santa Maria, RS, Brasil, gabriel.sampaio@eac.ufsm.br

Resumen

Generar una respuestas impulsiva biauricular de sala (RIBSs) de alta definición para una posición determinada requiere una cantidad significativa de recursos computacionales, siendo poco práctico para realidad virtual acústica. El principal problema son los miles de coeficientes que poseen las RIBSs de alta resolución, incluso en entornos de baja reverberación. En este sentido, este trabajo presenta una metodología para la generación de RIBSs de alta resolución a partir de información posicional del receptor acústico a un costo computacional adecuado para entornos en tiempo real utilizando un enfoque de aprendizaje profundo. La metodología consiste en: 1) generar un conjunto de datos con suficientes RIBSs de alta resolución distribuidos en un escenario virtual definido; 2) entrenar el autoencoder variacional para obtener RIBSs reducidas en dimensión a partir del codificador; 3) entrenar una red neuronal probabilística utilizando la posición del receptor acústico y la distancia euclidiana a la fuente de sonido como entradas; y las RIBSs reducidas como salidas; 4) aplicar el decodificador del autoencoder variacional para reconstruir RIBSs con la información espectral completa; 5) la RIBS reconstruida para la posición del receptor se convoluciona por segmentos con una señal sonora anecoica para producir la sensación acústica espacial. En este trabajo preliminar, se implementó la metodología hasta la obtención de las RIBSs reducidas al entrenar el autoencoder variacional. En aras de la viabilidad técnica y tecnológica, se utilizaron segmentos de respuestas impulsivas de sala (RISs) en lugar de RIBSs. Para evaluar el sistema se comparo el error cuadrático medio entre las RISs simuladas por un software acústico y las RISs generadas por el modelo. Finalmente, se realizó un análisis de costo computacional de la generación de RISs.

Palabras claves: Aurilización tiempo-real, Realidad Virtual Acústica, Variational Auto-encoders.

PACS: 43.55.Ka.

A deep learning approach for high resolution BRIR generation for real-time AVR.

Abstract

Generating a high-resolution binaural room impulse responses (BRIRs) for a given position requires a significant amount of computational resources, making it impractical for acoustic virtual reality in real-time environments. The main problem is the thousands of coefficients that high-resolution BRIRs possess, even in low-reverberant environments. In this sense, this work introduces a novel methodology for the generation of high-resolution BRIRs from receiver positional information with a computational cost suitable for real-time applications using a deep learning approach. The methodology consists of 1) generating a dataset with enough high-precision BRIRs distributed over a given scenario; 2) training variational autoencoder to obtain a dimensionally reduced BRIRs from the encoder; 3) training a neural network using the position of the acoustic receiver and the distance to the sound source as inputs and the reduced BRIRs as outputs; 4) apply the decoder of the variational autoencoder to reconstruct BRIRs with full spectral information; 5) the reconstructed BRIR for the receptor position is convolved by segments with an anechoic sound signal to produce the sensation of spatial acoustic effect. In this preliminary work, the methodology has been applied until the learning of reduced BRIRs. Also, segments of RIRs were used instead of the complete BRIRs for sake of technical viability. To test the results, the mean square error between the RIRs simulated by the acoustic software and the RIRs generated by the model were compared. Finally, computational cost to generate the RIR has been analyzed with respect to the total number of operations.

Keywords: Real-time Auralization, Acoustic Virtual Reality, Variational Auto-encoders.



1. INTRODUCCIÓN

La realidad virtual acústica (AVR, del inglés Acoustic Virtual Reality) se refiere al proceso de generar una sensación de inmersión acústica dentro de un ambiente dado, real o virtual, que el oyente experimentaría si estuviera presente en dicho ambiente [1]. Este proceso implica la utilización de algoritmos computacionales que, por un lado, modelan la propagación de onda del sonido y, por otro, implementan técnicas de aurilización [2].

Los modelos de propagación de la onda acústica para ambientes virtuales utilizan enfoques basados en simulaciones numéricas y modelamiento matemático, los cuales requieren un considerable esfuerzo computacional si se busca alcanzar representaciones realistas [3, 4]. Por otro lado, la aurilización es el núcleo de la AVR y consiste en la convolución entre cualquier señal sonora anecoica y la respuesta impulsiva binauricular de sala (BRIR, del inglés Binaural Room Impulse Response) correspondiente a la posición del oyente y orientación de la cabeza con respecto a la fuente sonora [2].

Las BRIRs pueden ser obtenidas experimentalmente o sintetizadas por simulación computacional. Las dificultades para la obtención experimental de las BRIRs en un ambiente real son el costo de los equipos de medición especializados, la necesidad de personal capacitado y el número de mediciones necesarias para representar adecuadamente todas las posibilidades acústicas dentro del ambiente en cuestión. En tal virtud, una alternativa es la generación de BRIRs por medio de simulación computacional.

La generación de BRIRs usando el método convencional implica un proceso de convolución en la que interviene la respuesta impulsiva relativa a la cabeza (HRIR, del inglés Head-Related Impulse Response o su transformada de Fourier, HRTF, del inglés Head-Related Transfer Function) y el frente de onda sonora que llega al receptor simulado por un rayo acústico [5]. El rayo acústico contiene la información relacionada con el espectro de frecuencia, la dirección de arribo representada por los ángulos de azimuth y elevación, y el tiempo de propagación. En [6] es descrito el procedimiento para generar BRIRs

que consiste en: 1) cada espectro de frecuencia del rayo acústico que arriva al receptor se multiplica con la HRTF correspondiente a la dirección de arribo; 2) se obtiene la HRIR filtrada al aplicar la transformada inversa de Fourier del resultado anterior; 3) cada HRIR filtrada es retrasada de acuerdo con el tiempo de propagación del rayo acústico incidente y; 4) se suman todas las HRIR retrasadas y filtradas que llegan al receptor acústico. Según [7], para una simulación fiable se requiere cerca de 10^5 rayos acústicos.

En ese sentido, las dificultades que enfrentan un abordaje de simulación computacional se relacionan al alto costo computacional asociado a la generación de las BRIRs de alta definición. La generación de BRIRs es aún más compleja cuando se simula el movimiento de un individuo dentro del entorno en tiempo real debido a que las BRIRs varían con respecto a la posición del receptor, siendo necesaria la generación de nuevas BRIRs por cada variación en la ubicación del oyente. Este constituye el principal cuello de botella para obtener una AVR confiable y en tiempo real.

En la última década, los enfoques basados en datos, como el aprendizaje de máquina (ML, del inglés Machine Learning) han sido una alternativa accesible gracias al enorme cantidad de información generada en varios campos de la industria e investigación, y la acústica no es la excepción. ML ha hecho un progreso significativo al proveer respuestas computacionales más eficientes respecto a técnicas convencionales de procesamiento de señales, principalmente en la etapa de ejecución. Sin embargo, requieren un largo proceso de entrenamiento y un conjunto de datos confiables y depurados [8].

La técnica más popular de ML son las redes neuronales artificiales (ANN, del inglés Artificial Neural Network) debido a su capacidad de solucionar problemas no lineales, generalización, tolerancia al ruido y bajo costo computacional en fase de ejecución [9]. Las ANNs y otros abordajes de ML han sido utilizados en la estimación de la HRIR del método convencional [9–12], alcanzando una reducción del costo computacional entre el 60 % y 90 %.

Sin embargo, a pesar de esta importante reduc-

ción computacional, no es suficiente para la generación de AVR de alta definición y en tiempo real porque la generación aún costosa de la BRIR se realiza solamente para una posición fija tanto para el oyente como para la fuente de sonido. Es decir, se debe generar una BRIR de alta definición cada vez que existe un cambio en la posición u orientación y esto implica repetir todo el proceso de generación de BRIRs desde el principio, lo cual hace que este enfoque sea computacionalmente costoso.

Los autores [13, 14] presentan enfoques basados en métodos de aprendizaje profundo para reconstruir el campo sonoro en ambiente cerrados. En [13] se reconstruye la presión del sonido de salas rectangulares en el ancho de banda de 30-300 Hz entrenando una red neuronal U-net usando datos simulados con pocas irregularidades en la sala dentro de una malla con micrófonos ubicados arbitrariamente. En el trabajo de [14] se plantea un enfoque de regresión de onda plana para la reconstrucción de las respuestas impulsivas de la sala mediante una red adversaria generativa (GAN, del inglés Generative Adversarial Network) para un recinto cerrado. Sin embargo, el enfoque asume que existen pocas ondas en el ambiente. Aunque estos trabajos son innovadores, la reconstrucción de BRIRs es simple e incompleta en el espectro de frecuencia.

Un problema adicional en la generación de BRIRs con toda la información espectral recae en su alta dimensionalidad temporal. Por ejemplo, una BRIR que tiene una duración de 2 segundos implica 88200 muestras, con una frecuencia de muestreo de 44.1 kHz, para cada oído. La alta dimensionalidad es un gran problema para cualquier técnica de aprendizaje de máquina ya que implica altos recursos computacionales en la fase de entrenamiento/aprendizaje.

Los abordajes de aprendizaje profundo, como *autoencoder* y redes generativas adversarias [15], son una alternativa interesante para abordar este problema ya que realizan una extracción automática de representaciones comprimidas del *raw data* con el objetivo de reconstruir completamente la información original a partir de estas representaciones latentes, [15, 16]. En tal virtud, el objetivo del presente trabajo es proponer una

nueva metodología para la generación de BRIRs de alta definición en tiempo real usando técnicas de aprendizaje profundo como los *autoencoders varacionales* (VAE, del inglés Variational Autoencoders).

Este trabajo está organizado de la siguiente manera. En la Sección 2 se explica el fundamento matemático de los VAEs. En la Sección 3 se detalla la metodología propuesta para la generación de BRIRs de alta definición para aplicaciones en tiempo real basados en aprendizaje profundo. Adicionalmente, se explica el procedimiento realizado para la evaluación del primer sistema de la metodología propuesta que consiste en la generación de BRIRs usando un VAE. Cabe mencionar que, por simplicidad en la implementación del VAE y también con el objetivo de evaluar la precisión de los resultados, se usaron las respuestas impulsivas de la sala (RIRs, del inglés Room Impulse Responses) en lugar de BRIRs. En la Sección 4 se muestran los resultados de la generación de RIRs a partir de una dimensión reducida. Finalmente, en la Sección 5 se presenta las conclusiones y análisis del costo computacional del decodificador del VAE.

2. MARCO TEÓRICO

En esta parte del trabajo se explica algunos conceptos matemáticos necesarios del VAE que se encuentran en el trabajo de Kinga y Welling, [15]. En esta sección $\mathbf{x} = [x_1, x_2, \dots, x_n]$, se refiere al vector de observaciones, mientras que $\mathbf{z} = [z_1, z_2, \dots, z_d]$ al vector de variables aleatorias latentes.

Los autoencoders son técnicas de aprendizaje de máquina auto supervisados basados en ANNs con el objetivo de inferir una dimension reducida de las observaciones $\mathbf{x}$, a este proceso en ML se le conoce como extracción de características. En los autoencoders esta dimensión reducida también se la denomina como espacio latente u oculto. La idea general de los autoencoders consiste en configurar un codificador y un decodificador usando redes neuronales y aprender el mejor esquema de codificación-decodificación mediante un proceso de optimización iterativo. Es decir, se obtiene un vector del codificador donde cada dimensión representa un atributo aprendido



a partir de las observaciones. Cada dimensión del espacio latente es representado por un valor real. Una vez aprendido el espacio latente por el codificador, se puede obtener nuevamente la observación, $\mathbf{x}$, usando el decodificador del autoencoder.

Por otro lado, los VAEs proporcionan un enfoque probabilístico para describir una observación en el espacio latente. Es decir, en lugar de construir un codificador que genere valor único para representar cada atributo del estado latente, el codificador del VAE aprende una distribución de probabilidad para cada atributo latente. Sin embargo, ya que los modelos de probabilidad son inadecuados para problemas complejos con observaciones de alta dimensionalidad, Kinga y Welling [15], introducen un modelo de probabilidad basado en ANNs que permiten solventar estas limitaciones. Esto se debe a que los abordajes basados ANNs son escalables computacionalmente gracias a algoritmos de optimización eficientes.

El VAE es un modelo de probabilidad, $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ que depende de las observaciones, $\mathbf{x}$, y las variables latentes, $\mathbf{z}$. El objetivo del *modelo generativo o decodificador* del VAE, $p_\theta(\mathbf{z}|\mathbf{x})$, es generar una observación, $\mathbf{x}$, dado las variables latentes, $\mathbf{z}$. Mientras que el objetivo del *codificador o aproximación a posterior*, $p(\mathbf{x}|\mathbf{z})$, es inferir las mejores variables latentes, $\mathbf{z}$, a partir de las observaciones $\mathbf{x}$. Sin embargo, derivar $p(\mathbf{z}|\mathbf{x})$ es complejo ya que usualmente es indefinida desconocida. No obstante, $p(\mathbf{z}|\mathbf{x})$ se puede estimar mediante técnicas estadísticas pero con un alto costo computacional. Por esta razón, Kinga y Welling [15] aplican inferencia variacional para aproximar $p(\mathbf{z}|\mathbf{x})$ a un costo computacional menor.

Según [17], la inferencia variacional consiste en aproximar la densidad a posterior desconocida, $p(\mathbf{z}|\mathbf{x})$, a una densidad posterior variacional de un conjunto de densidades conocidas y definidas por los parámetros ϕ . El objetivo es encontrar un miembro de esta familia que minimice la divergencia Kullback-Liebler (KL) hasta obtener la distribución a posterior real. En conclusión, se debe encontrar los parámetros de ϕ que minimi-

cen la divergencia KL, definida por

$$q_\phi(\mathbf{z}|\mathbf{x}) = \arg \min_{\phi} \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})). \quad (1)$$

La divergencia KL, Ecuación (2), mide la información perdida al usar q_ϕ para aproximar p_θ , siendo un término no negativo y es cero cuando $q_\phi = p_\theta$.

$$\text{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})] = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})} \right]. \quad (2)$$

Debido a que no se puede derivar la divergencia KL, ya que $p_\theta(\mathbf{z}|\mathbf{x})$ es desconocida, es necesario definir una función objetivo alternativa para poder actualizar los parámetros θ y ϕ , [15, 17]. Para lo cual, primero se expande la Ecuación (2) y luego se aplica la esperanza/promedio con respecto a $q_\phi(\mathbf{z}|\mathbf{x})$, obteniendo

$$\begin{aligned} \text{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})] &= \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{x})}{p_\theta(\mathbf{x}, \mathbf{z})} \right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{x}, \mathbf{z})} \right] + \log p_\theta(\mathbf{x}). \end{aligned} \quad (3)$$

De la Ecuación (3) se puede observar que la divergencia KL depende de $\log p_\theta(\mathbf{x})$ y la nueva función objetivo llamada *Evidence Lower Bound* (ELBO, por su sigla en inglés), definida por:

$$\mathcal{L}_{\theta, \phi}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})]. \quad (4)$$

Según la Ecuación (2) la divergencia KL es no negativa. Por lo tanto, ELBO es el límite inferior de log-likelihood de las observaciones $\mathbf{x}$, tal como se muestra en:

$$\begin{aligned} \mathcal{L}_{\theta, \phi}(\mathbf{x}) &= \log p_\theta(\mathbf{x}) - \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})); \\ \mathcal{L}_{\theta, \phi}(\mathbf{x}) &\leq \log p_\theta(\mathbf{x}). \end{aligned} \quad (5)$$

Según [15], al maximizar $\mathcal{L}_{\theta, \phi}(\mathbf{x})$ con respecto a los parámetros, θ y ϕ , se logra al mismo tiempo: 1) maximizar la densidad marginal, $p_\theta(\mathbf{x})$ (esto significa que el modelo generativo o decodificador reduce el error en reconstruir $\mathbf{x}$); 2) minimizar la divergencia KL entre la distribución a posterior aproximada, $q_\phi(\mathbf{z}|\mathbf{x})$, y la a posterior real, $p_\theta(\mathbf{z}|\mathbf{x})$, por lo tanto, el modelo del codifi-

cador, q_ϕ , es mejor.

Dado un conjunto de datos distribuidos idéntica e independientemente, la función objetivo del ELBO es la suma o promedio de ELBO por cada observación [15]. Es decir

$$\mathcal{L}_{\theta,\phi}(\mathcal{D}) = \frac{1}{N} \sum_{x \in \mathcal{D}} \mathcal{L}_{\theta,\phi}(x). \quad (6)$$

Donde N es el número de observaciones del conjunto de datos $\mathcal{D}$, y x una observación. Sin embargo, tanto $\mathcal{L}_{\theta,\phi}(x)$ como su gradiente, $\nabla_{\theta,\phi} \mathcal{L}_{\theta,\phi}(x)$, son indefinidos. En su lugar, una estimación Monte Carlo puede aplicarse para estimar ELBO. En [15] se introduce un estimador para ELBO que usa una muestra $\{z_i\}_{i=1}^L$ de $q_\phi(z|x)$, definido por

$$\hat{\mathcal{L}}^A(\theta, \phi; x) = \frac{1}{L} \sum_{j=1}^L \log p_\theta(x, z_j) - \log q_\theta(z_j|x). \quad (7)$$

En principio, tanto θ y ϕ pueden ser actualizados mediante el algoritmo de descenso del gradiente estocástico (SGD, del inglés Stochastic Gradient Descent), aplicando la derivada de cada parámetro con respecto a $\mathcal{L}$ por cada observación. La derivada con respecto a θ es calculada fácilmente por SGD. Sin embargo, los valores z_j no son funciones diferenciables respecto a ϕ , debido a que son muestras de la distribución $q_\theta(z_j|x)$. Para solventar este problema se usa la técnica de reparametrización [17].

La técnica de la reparametrización permite que la derivada con respecto al parámetro ϕ de la distribución de la función objetivo se retropropague por la ANN. Esto se logra al cambiar la forma de tomar las muestras de la distribución $q_\theta(z_j|x)$. En vez, de obtener la muestra directamente de $q_\theta(z|x)$, se usa una variable auxiliar ϵ_j de la distribución $p(\epsilon)$ que no está parametrizada por ϕ . Por tanto, el primer paso es escoger una distribución, $p(\epsilon)$, y una transformación $g_\phi(\epsilon, x)$, tal que: $q_\theta(z|x)$, tenga la misma distribución que $g_\phi(\epsilon, x)$, donde $\epsilon \sim p(\epsilon)$. En resumen, el proceso de muestreo usando la transformación es equivalente a muestrear de $q_\theta(z|x)$. Por simplicidad, se asume que $g_\phi(\epsilon, x)$ y $p(\epsilon)$ son conocidas y que satisfacen los criterios anteriores.

Por tanto, la Equación (7) ya es diferenciable

con respecto a ϕ y θ , es decir, los parámetros ya pueden actualizarse usando el algoritmo SGD. Donde $z_j = g_\phi(\epsilon, x)$; y ϵ_j ha sido muestreada de $p(\epsilon)$. En la Figura 1 se presenta la arquitectura global del VAE en la que se detalla la técnica de la reparametrización.

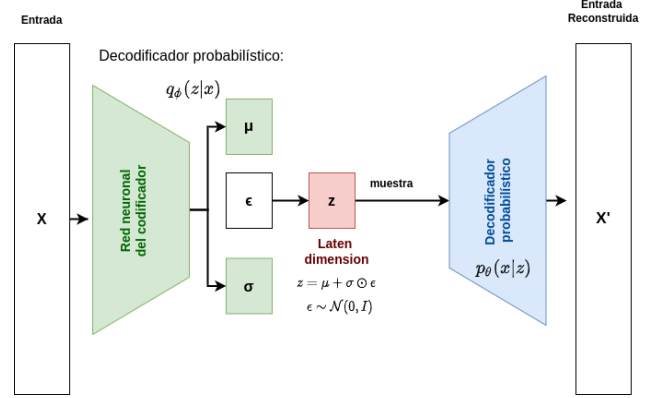


Figura 1: Arquitectura general del VAE. Fuente [18]

3. METODOLOGÍA

Esta sección detalla la metodología propuesta para la generación de BRIRs de alta fidelidad para ambientes en tiempo real. La metodología consta de tres sistemas:

- A Adquisición de datos, agrupación y extracción de características;
- B Generación de BRIRs reducidas a partir de posición del receptor acústico;
- C Generación de AVR en tiempo real.

El primer sistema consiste en: 1) realizar un proceso de adquisición de datos a partir de un número representativo de BRIRs de alta fidelidad generadas por un simulador acústico, correspondientes a diferentes posiciones de los oyentes dentro de la sala distribuidas homogéneamente; 2) con el objetivo de contar con el número suficiente de BRIRs para el entrenamiento de sistemas de aprendizaje de máquina se aplica un proceso de “data augmentation” mediante interpolación bilineal de las BRIRs obtenidas del simulador acústico. Las BRIRs resultantes del proceso de interpolación también son distribuidas homogéneamente dentro del área que encierran las BRIRs simuladas; 3) realizar un proceso de agrupación usando las posiciones de los receptores con el objetivo de obtener conjuntos



Figura 2: Metodología propuesta para la generación de BRIRs de alta fidelidad y en tiempo real. Fuente: Autores.

más homogéneos y, por ende, minimizar la dificultad de modelamiento de las funciones; 4) extraer características reducidas de las BRIRs por medio de un VAE por cada grupo de receptores acústicos definidos en el paso anterior. Como fue señalado en la sección previa, el codificador del VAE aprende una versión de dimensión reducida de la BRIR (BRIR codificada), mientras que el decodificador nos permite reconstruir las BRIRs de alta definición a partir de la dimensión reducida.

El segundo sistema “generación de BRIRs reducidas a partir de posición del receptor acústico” está compuesta por 1) preparar un data set que contiene: la posición del receptor, la orientación de cada receptor y la respectiva BRIRs de dimensión reducida por cada clúster; 2) aprender las BRIRs de dimensión reducida a partir de la información posicional al entrenar un modelo de probabilidad basado en ANNs. La idea es ajustar un modelo de probabilidad basado ANNs a la distribución de probabilidad de las BRIRs reducidas aprendida por el codificador del AVE por cada clúster.

El tercer sistema “generación de realidad virtual

acústica en tiempo real” tiene dos fases: 1) obtener la BRIR de alta definición que consiste en: a) determinar el clúster a cual pertenece la BRIRs a partir de la información posicional del receptor; b) selecciona tanto el modelo de probabilidad basado en ANNs y el decodificador VAE correspondiente al clúster; c) generar la BIRS de alta definición al ejecutar el modelo de probabilidad y luego el decodificador del VAE determinado anteriormente. En resumen, esta fase devuelve una BRIR de alta definición a partir posición del receptor a un costo computacional menor, gracias a la reducción de la dimensionalidad. 2) realizar un proceso de convolución segmentada entre la BRIR resultante y cualquier señal de audio anecoico para sintetizar un sonido que genere la sensación de inmersión acústica al oyente. En la Figura 2 se muestra la metodología y los sistemas que la componen. Cabe mencionar que todo el proceso para la generación de una BRIR se debe repetir cada vez que exista un cambio posicional o de orientación del receptor acústico.

Este trabajo se centra en detallar la implementación del sistema de adquisición, agrupación y extracción de características. En este sentido, debido a la limitación de los recursos computacio-

nales y con el objetivo de garantizar la viabilidad técnica al momento de implementar el VAE se decidió reducir la complejidad del problema. Para lo cual se realizó lo siguiente: 1) solo se usó las RIRs generadas por el simulador acústico en vez de las BRIRs; 2) debido a que el tiempo de reverberación de la sala virtual seleccionada es aproximadamente de 2 segundos, esto implica que existen aproximadamente 88200 muestras por cada RIR a una frecuencia de muestreo de 44.1 KHz. Esta alta dimensionalidad de la RIR demanda considerables recursos computacionales para implentar el VAE, por tanto, se dividió las RIRS a segmentos de 0.3 segundos, es decir, 13017 muestras que permitieron el entrenamiento del VAE usando los recursos computacionales disponibles en el Laboratorio de Simulación Numérica y Acústica (LASINAC) de la Escuela Politécnica Nacional (E.P.N) en Quito-Ecuador.

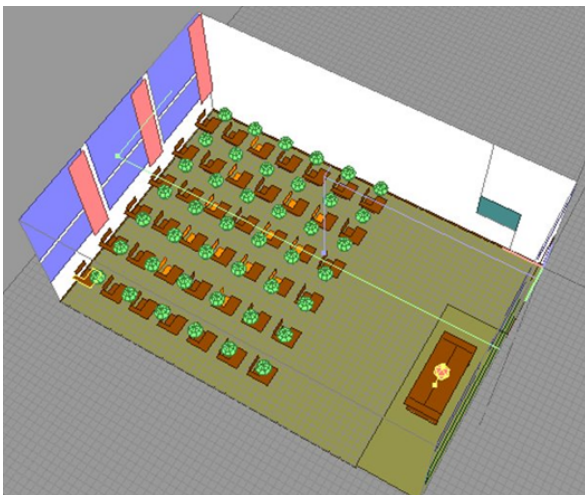


Figura 3: Ubicación de los 42 receptores acústicos en una sala virtual (aula de clase) generada por el simulador acústico RAIOS 7 representadas por las esferas verdes. La esfera de color rojo representa la fuente sonora. Cada receptor acústico genera una RIR. Figura tomada del software acústico RAIOS 7, [19]

El primer paso del sistema de adquisición, agrupación y extracción de características es obtener las RIRs del simulador acústico. Para lo cual, se definió primero el escenario virtual. Se seleccionó a una aula de clase que contiene 42 receptores acústicos ubicados donde usualmente se encuentran los oyentes, y una fuente sonora ubicada en donde se encuentra el locutor, tal como se lo muestra en la Figura 3. Una vez definido el escenario, se simuló 42 RIRs en el software acústico RAIOS 7 [19]. A partir de estas 42

RIRs se aplicó el proceso de data augmentation, obteniendo un data set de alrededor de 120.000 RIRs de alta definición distribuidas uniformemente dentro del área donde se encuentran las RIRs simuladas. En la Figura 4 se muestra la distribución espacial de las posiciones de las RIRs interpoladas en la sala virtual representadas por el área de color púrpura.

El proceso de agrupación se realizó usando la técnica Gaussian Mixture Model que tomó como características para el agrupamiento las posiciones de los receptores acústicos. A través de un proceso de prueba y error se definió el número de clústers en 9, tal como se observa en la Figura 5. Cada RIRs del dataset se asignó a un clúster y del conjunto de RIRs resultante por clúster el 70 % se asignó para el entrenamiento, el 15 % para la evaluación y el 15 % para la validación del VAE.

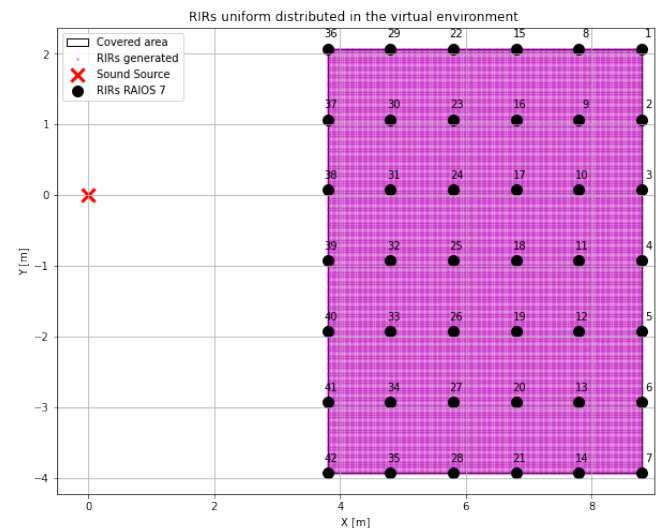


Figura 4: Representación espacial de las posiciones de las RIRs interpoladas a partir de 42 RIRs generadas por el simulador acústico. Los puntos púrpuras representan las posiciones donde las RIRs fueron interpoladas. Los círculos negros representan la posición de las 42 RIRs simuladas. La cruz representa la ubicación de la fuente de sonido. Fuente: Autores.

La arquitectura del VAE fue definida de la siguiente manera. El codificador consta de 3 capas, la capa de entrada tiene 13107 neuronas correspondiente a la dimensionalidad de la RIR, la capa intermedia consta de entre el 20 % y el 30 % de las neuronas de entrada. Finalmente, la capa de salida devuelve una distribución gaussiana multivariable entre 131 y 260 variables correspondientes al 1 % y 2 % de la dimensión de

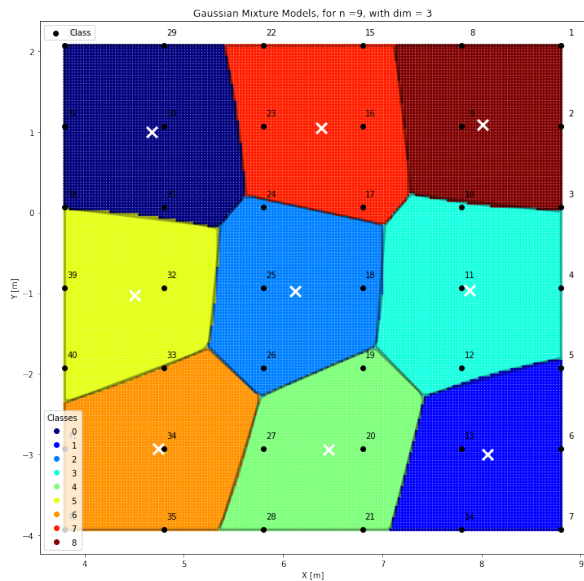


Figura 5: Clusterización en base a la posición del receptor acústico y la distancia euclidiana con respecto a la fuente sonora. Cada área colorida representan un clúster, los puntos negros la posición de los receptores acústicos en la sala virtual. Fuente: Autores

entrada. El decodificador tiene como capa de entrada la distribución de probabilidad mencionada previamente, la capa intermedia es igual a la del codificador y la capa de salida corresponde a una capa de probabilidad con 13107 distribuciones gaussianas independientes.

Sin embargo, la reducción de dimensionalidad no es el único reto de la metodología propuesta. Otra dificultad está en el sistema de generación de BRIRs debido a que no se conoce con certeza si existe una relación causal significativa entre la posición del receptor y la BRIR codificada para reconstruir un segmento de dicha BRIR con toda la información espectral a un error aceptable y con costo computacional reducido que permita un procesamiento en tiempo real. Esta problemática se abordará en el sistema de generación de BRIRs de alta definición que se encuentra en la fase de implementación.

En resumen, aunque la metodología propuesta está en una etapa inicial y tiene aún un camino por recorrer, es novedosa ya que en la revisión de la literatura no se encontró un enfoque que genere BRIRs de alta definición usando únicamente la información posicional del receptor.

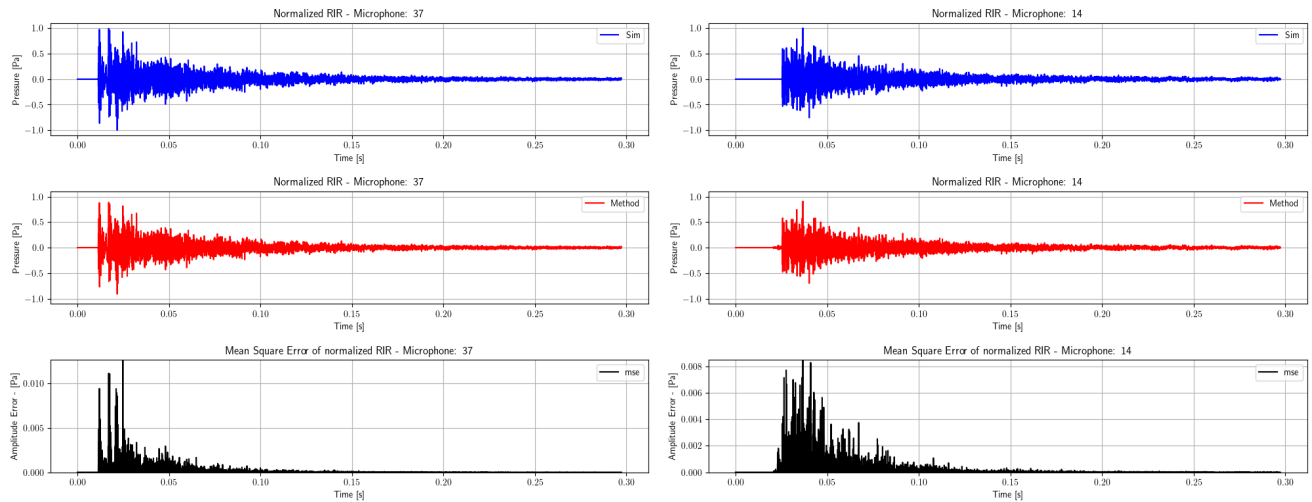
4. RESULTADOS

En esta sección se presentan y analizan los resultados del sistema de adquisición de datos, agrupación y extracción de características de la metodología propuesta. Se realizó el entrenamiento de un VAE por cada clúster variando los siguientes hiperparámetros: el algoritmo de optimización; el número de neuronas por capa; y las funciones de activación. La mejor arquitectura en cada clúster se obtuvo con: 3932 neuronas en la capa intermedia, 157 variables en la *latent dimension*, la función de activación sigmoide en ambas capas, y el optimizador RMSprop con tasa de aprendizaje de 5×10^{-4} y $\rho = 0,8$.

Para evaluar el sistema de adquisición de datos, agrupación y extracción de características se comparó las 42 RIRs generadas por el software de simulación acústica con las RIRs correspondientes reconstruidas por el VAE en el dominio del tiempo.

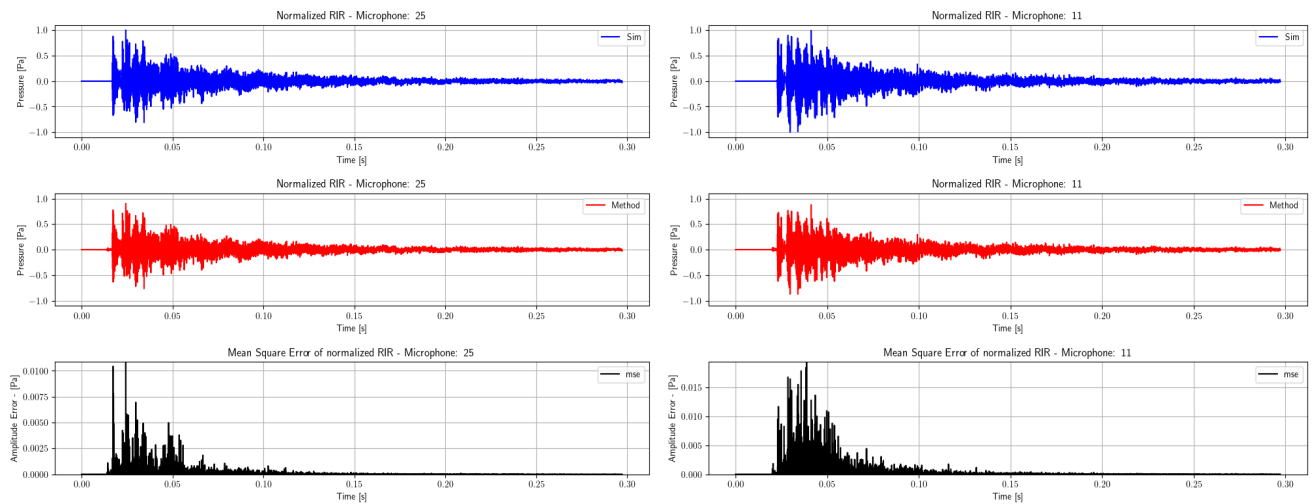
Cada subfigura de la Figura 6 corresponde a un clúster. En donde: en la parte superior se observa la RIR normalizada simulada por RAIOS 7, en azul; en la mitad la RIR normalizada generada por el VAE, en rojo; y en la parte inferior se presenta el error cuadrático medio entre la RIRs simulada y la generada.

Para evaluar si la generación de BRIRs es adecuada para aplicaciones en tiempo real se midió el tiempo computacional que toma sintetizar cada una de las 42 RIRs de prueba utilizando el VAE. El tiempo promedio para la generación de una fracción de 297.21 ms de una RIR es de alrededor de 31 ms. En [1], los autores consideran a la latencia máxima como medida de rendimiento para entornos virtuales auditivos dinámicos. La latencia total de un sistema de entornos virtuales acústicos se define como el tiempo transcurrido desde que sucede un evento, como la rotación de la cabeza y/o el cambio de posición del receptor, hasta que las consecuencias de este evento provocan cambios significativos en el entorno virtual percibido por el sonido aurilizado del sistema. La latencia total del sistema, por tanto, toma en consideración varios factores entre ellos los algoritmos aurilización. Lindau y Weinzier [20] resaltan que para una sintetización plausible de aurilización la latencia no debe exceder los



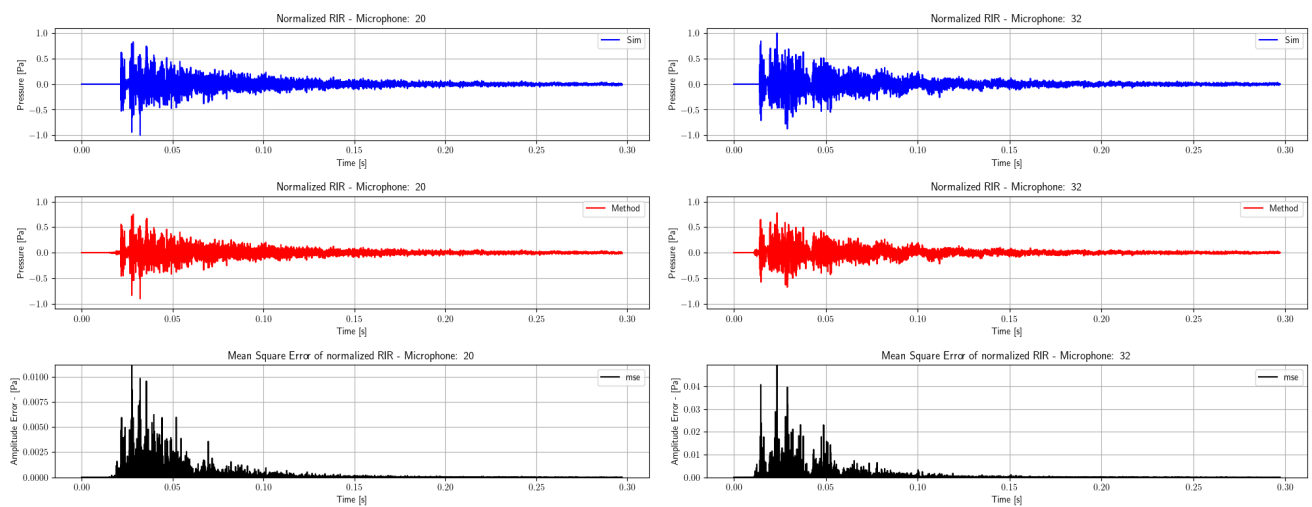
(a) RIR normalizada del receptor 37 del clúster No. 0.

(b) RIR normalizada del receptor 14 del clúster No. 1.



(c) RIR normalizada del receptor 25 del clúster No. 2.

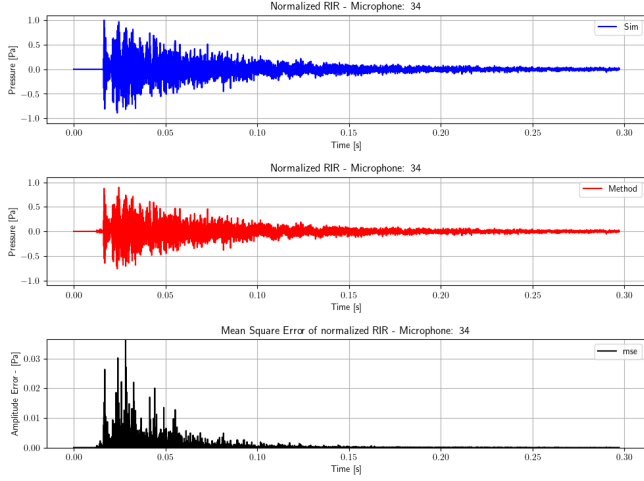
(d) RIR normalizada del receptor 11 del clúster No. 3.



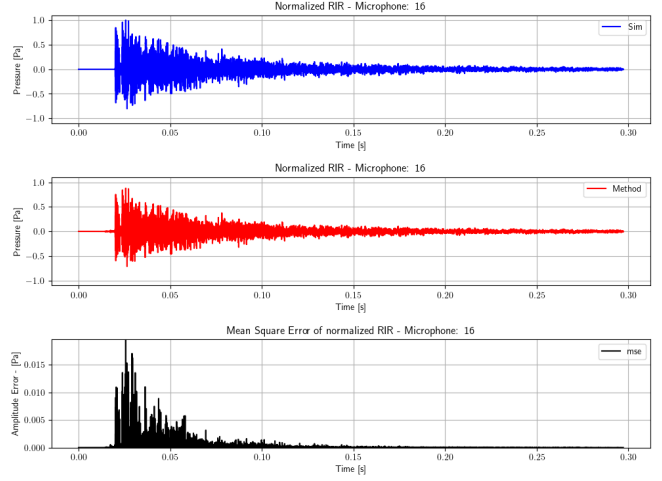
(e) RIR normalizada del receptor 20 del clúster No. 4.

(f) RIR normalizada del receptor 32 del clúster No. 5.

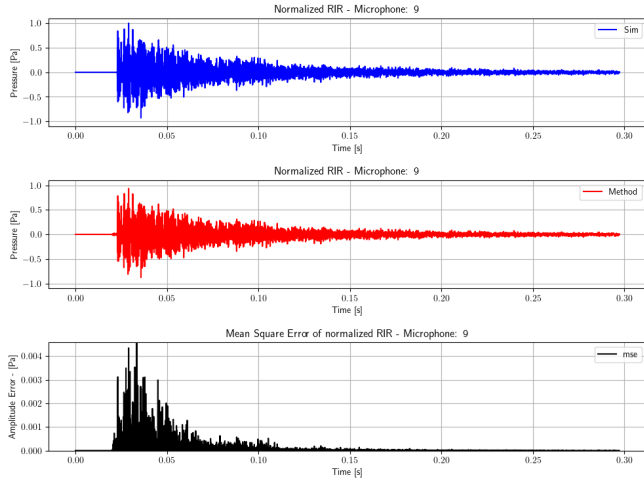
Figura 6: Parte 1.



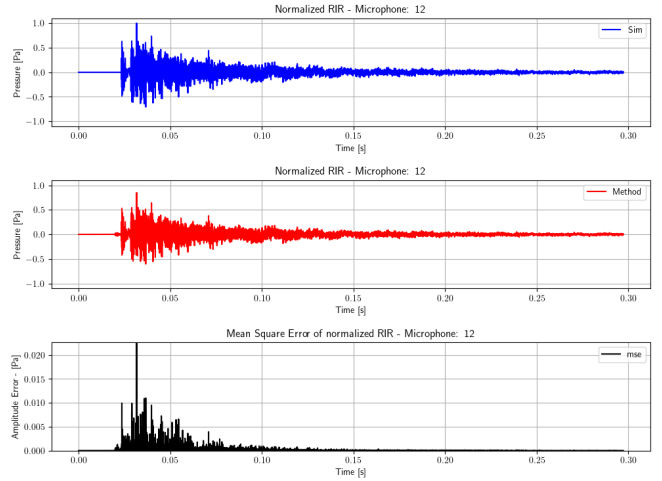
(g) RIR normalizada del receptor 34 del clúster No. 6



(h) RIR normalizada del receptor 16 del clúster No. 7



(i) RIR normalizada del receptor 9 del clúster No. 8



(j) RIR normalizada del receptor 12 del clúster No. 3

Figura 6: Parte 2. En la parte superior de cada subfigura se muestra la RIRs normalizada simulada por RAIOS 7, en azul; en el medio la RIR normalizada generada por el VAE, en rojo. En la parte inferior, se muestra el error cuadrático medio de las RIRs normalizadas en negro. Fuente: Autores.

112 ms. En otros trabajos como [20–22] proponen un rango para el umbral aceptable para la latencia entre 52 ms y 85 ms. Otro indicador a considerar es el efecto de precedencia o efecto Haa que enuncia: si un sonido es seguido por otro sonido separado por un retraso por debajo del umbral del eco del oyente 50 ms, el oyente percibe un solo evento auditivo.

5. ANÁLISIS Y CONCLUSIONES

5.1 Análisis del costo Computacional

En [9] se realiza un análisis del costo computacional del modelo basado en ANNs para la generación de HRIR modificadas en función del

total de operaciones requeridas. En este sentido, para una red neuronal de n capas el total de operaciones, que depende del número de parámetros (i.e. pesos), está definido por

$$T_{0ANN} = 2 \left[\sum_{i=1}^{n-1} (L_i + 1) L_{i+1} \right]. \quad (8)$$

Donde L_i es la dimensión del vector parámetros de la i -ésima capa, entendiendo que cuando $i = 1$ se refiere a la capa de entrada, cuando $i = n$ se habla de la capa de salida y los valores intermedios se refieren a las capas ocultas. T_{0ANN} se refiere al total de operaciones del modelo basado en ANN.

El costo computacional del modelo propuesto de generación de BRIRs recae en su mayoría en la ejecución del decodificador del VAE. Por tanto, se realiza un análisis del costo computacional del mismo, el cual (como mencionado en la Sección 2) es un modelo de probabilidad basado en ANNs. En la arquitectura del VAE, en lo que se refiere al decodificador, el modelo de probabilidad $p_{\theta}(\mathbf{x}|\mathbf{z})$ se encuentra en la capa de salida de la ANN. Donde $p_{\theta}(\mathbf{x}|\mathbf{z})$ está compuesta de funciones de distribución Gaussianas independientes $\mathcal{N}(\mu_i, \sigma_i)$. Por tanto, el total de operaciones del decodificador es

$$TO_{DE} = 2 \left[\sum_{i=1}^{m-2} (L_i + 1)L_{i+1} + (L_{m-1} + 1)(2L_m) \right]. \quad (9)$$

En la Ecuación (9), m se refiere al número de capas de la ANN del decodificador y TO_{DE} al total de operaciones del decodificador. Entonces, para $m = 3$ capas del decodificador del VAE el total de operaciones es

$$TO_{DE} = 2 [(158)(3932) + (3933)(2 \times 13107)] \\ = 207441836.$$

Cabe recordar que, para la generación de una BRIR de alta resolución usando la metodología propuesta se requiere el costo computacional del decodificador del VAE y del modelo de probabilidad basado en ANN para obtener las BRIRs de dimensión reducida a partir de la posición (la cual será contemplada posteriormente en la metodología propuesta).

5.2 Conclusiones

Se presentó una nueva metodología para la generación de BRIRs de alta definición para entornos en tiempo real usando VAE. Adicionalmente, se implementó el sistema adquisición de datos, agrupación y extracción de características usando solamente fragmentos de RIR de un conjunto de RIRs en lugar de BRIRs, obteniendo resultados alentadores.

Las RIRs correspondientes a los receptores acústicos que se ubican cerca del centro de cada clúster tiene un error cuadrático medio menor, mientras que este aumenta cada vez que las RIRs correspondientes a las posiciones de los recep-

tores acústicos se acercan al borde del clúster, como era de esperarse. Además, las RIRs de los receptores acústicos que tienen un mayor error cuadrático medio son aquellos que se encuentran cerca de las esquinas y cerca de las paredes del entorno virtual, ver en la Figura 3. Esto se debe a que, en estas posiciones las RIRs simuladas tienen un comportamiento atípico comparado a las que se encuentran en otras posiciones debido a que existen más reflexiones de las ondas sonoras que afectan principalmente a la parte inicial de las RIRs. Según la ISO 3382-1, las mediciones de las BRIRs y RIRs para una sala deben realizarse al menos a 1.5 metros de distancia de las paredes o fuentes sonoras.

Los principales problemas en la generación de BRIRs de alta definición y en tiempo real son: la alta dimensionalidad y el costo computacional. Esto dificulta la implementación, tanto en el entrenamiento como en la ejecución, de cualquier modelo basado en aprendizaje de máquina ya que los recursos computacionales para este fin son altos. Debido a lo expuesto, este trabajo se enfocó en la reconstrucción de solo un fragmento de las RIRs a partir de una su dimensión reducida. No obstante, esto no representa un problema dado que la generación del audio depende de un proceso de convolución segmentada, para lo cual solo se necesita un fragmento de la señal.

Aunque los resultados hasta el momento parecen alentadores, se debe considerar para trabajos futuros entrenar el VAE con otros tipos de ANNs en la capas intermedias (e.g. LSTM-ANNs, del inglés Long-Short Term Memory-ANNs). Es importante mencionar que tanto el sistema de generación de BRIRs de alta definición como el sistema de generación de AVR se encuentra en etapa de investigación y desarrollo. Finalmente, aún es necesario realizar pruebas psicoacústicas de las BRIR or RIRs generadas por el método para validarlas.

REFERENCIAS

- [1] Tommasini, Fabián C.; Ramos, Oscar A.; Hüg, Mercedes X. e Ferreyra, Sebastián P. A Computational Model to Implement Binaural Synthesis in a Hard Real-Time Auditory Virtual Environment. *Acoust. Aust.*, 47(1):51–66, 2019. ISSN 18392571. doi:[10.1007/s40857-019-00152-7](https://doi.org/10.1007/s40857-019-00152-7). Disponible en:



- <https://doi.org/10.1007/s40857-019-00152-7>.
- [2] Vorländer, Michael. Virtual Acoustics. 39(3):307–318, 2014. ISSN 01375075. doi:[10.2478/aoa-2014-0036](https://doi.org/10.2478/aoa-2014-0036).
 - [3] Schröder, Dirk; Pelzer, Sönke; Vetterli, Martin e Vorländer, Michael. Through the hourglass: A faithful audiovisual reconstruction of the old montreux casino. *Acoust. Aust.*, 43(1):49–57, 2015. ISSN 08146039. doi:[10.1007/s40857-015-0009-4](https://doi.org/10.1007/s40857-015-0009-4). Disponível em: <http://dx.doi.org/10.1007/s40857-015-0009-4>.
 - [4] Pulkki, Ville e Svensson, U. Peter. Machine-learning-based estimation and rendering of scattering in virtual reality. *J. Acoust. Soc. Am.*, 145(4):2664–2676, 2019. ISSN 0001-4966. doi:[10.1121/1.5095875](https://doi.org/10.1121/1.5095875). Disponível em: <http://dx.doi.org/10.1121/1.5095875>.
 - [5] Vorländer, Michael e Summers, Jason E. Auralization: Fundamentals of acoustics, modelling, simulation, algorithms, and facoustic virtual reality. *J. Acoust. Soc. Am.*, 123(6):4028, 2008.
 - [6] Tenenbaum, R A; Taminato, F O; Melo, V S G e Torres, J C B. Auralization generated by modeling HRIRs with artificial neural networks and its validation using articulation tests. *Appl. Acoust.*, 130:260–269, 2018. doi:[10.1016/j.apacoust.2017.09.025](https://doi.org/10.1016/j.apacoust.2017.09.025).
 - [7] Bork, Ingolf. A comparison of room simulation software-the 2nd round robin on room acoustical computer simulation. *Acta Acustica united with Acustica*, 86(6):943–956, 2000.
 - [8] Bianco, Michael J; Gerstoft, Peter; Traer, James; Ozanich, Emma; Roch, Marie A; Gannot, Sharon e Deledalle, Charles-Alban. Machine learning in acoustics: Theory and applications. *J. Acoust. Soc. Am.*, 146(5):3590–3628, 2019.
 - [9] Naranjo, José F Lucio e Tenenbaum, Roberto A. *Inteligência computacional aplicada na geração de respostas impulsivas bi-auriculares e em aurilização de salas*. Tese de doutorado, Universidade do Estado do Rio de Janeiro, 2014.
 - [10] Taminato, Filipe Otsuka. *Redes neurais artificiais aplicadas à modelagem de respostas impulsivas associadas à cabeça humana para gerar aurilização*. Tese de doutorado, Universidade do Estado do Rio de Janeiro, 2018.
 - [11] Tommasini, Fabián C.; Ramos, Oscar A.; Hüg, Mercedes X. e Bermejo, Fernando. Usage of spectral distortion for objective evaluation of personalized HRTF in the median plane. *International Journal of Acoustics and Vibrations*, 20(2):81–89, 2015. ISSN 10275851. doi:[10.20855/ijav.2015.20.2371](https://doi.org/10.20855/ijav.2015.20.2371).
 - [12] Tamulionis, Mantas e Serackis, Arturas. Comparison of Multi-Layer Perceptron and Cascade Feed-Forward Neural Network for Head-Related Transfer Function Interpolation. *2019 Open Conference of Electrical, Electronic and Information Sciences, eStream 2019 - Proceedings*, págs. 1–4, 2019. doi:[10.1109/eStream.2019.8732158](https://doi.org/10.1109/eStream.2019.8732158).
 - [13] Lluís, Francesc; Martínez-Nuevo, Pablo; Bo Møller, Martin e Ewan Shepstone, Sven. Sound field reconstruction in rooms: Inpainting meets super-resolution. *J. Acoust. Soc. Am.*, 148(2):649–659, 2020.
 - [14] Karakostas, Xenofon e Fernandez Grande, Efrén. Sound field reconstruction in rooms with deep generative models. Em *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 263, págs. 1527–1538. Institute of Noise Control Engineering, 2021.
 - [15] Kingma, Diederik P e Welling, Max. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.
 - [16] Blundell, Charles; Cornebise, Julien; Kavukcuoglu, Koray e Wierstra, Daan. Weight uncertainty in neural network. Em *International conference on machine learning*, págs. 1613–1622. PMLR, 2015.
 - [17] Blei, David M; Kucukelbir, Alp e McAuliffe, Jon D. Variational inference: A review for statisticians. *J Am Stat Assoc*, 112(518):859–877, 2017.
 - [18] Sinitambirivoutin, Emrick. An introduction to Variational Auto Encoders (VAEs). <https://towardsdatascience.com/an-introduction-to-variational-auto-encoders-vaes-803ddfb623df>, 2020. [Online; accessed 8-Abril-2022].
 - [19] Tenenbaum, Roberto A.; Taminato, Filipe O. e Melo, Viviane S.G. Fast auralization using radial basis functions type of artificial neural network techniques. *Appl. Acoust.*, 157:106993, 2020. ISSN 0003682X. doi:[10.1016/j.apacoust.2019.07.041](https://doi.org/10.1016/j.apacoust.2019.07.041). Disponível em: <https://doi.org/10.1016/j.apacoust.2019.07.041>.
 - [20] Lindau, Alexander. The perception of system latency in dynamic binaural synthesis. *Proc. of 35th DAGA*, págs. 1063–1066, 2009.
 - [21] Mackensen, Philip. Auditive localization. head movements, an additional cue in localization. *PhD dissertation, Tech. Univ.*, 2004.
 - [22] Yairi, Satoshi; Iwaya, Yukio e Suzuki, Yôiti. Investigation of system latency detection threshold of virtual auditory display. Georgia Institute of Technology, 2006.