Universitat d'Alacant
Universidad de Alicante

Memorias del Programa de Redes-I3CE de calidad, innovación e investigación en docencia universitaria

Convocatoria
2020-21

Memòries del Programa de Xarxes-I3CE de qualitat, innovació i investigació en docència universitària

Convocatòria
2020-21

Satorre Cuerda, Rosana (Coordinación)
Menargues Marcilla, María Asunción; Díez Ros, Rocío; Pellín Buades, Neus (Eds.)

UA
UNIVERSITAT D'ALACANT
UNIVERSIDAD DE ALICANTE
Vicerectorat de Transformació Digital
Vicerrectorado de Transformación Digital
Institut de Ciències de l'Educació
Instituto de Ciencias de la Educación

Memorias del Programa de Redes-I3CE de calidad, innovación e investigación en docencia universitaria. Convocatoria 2020-21 / Memòries del Programa de Xarxes-I3CE de qualitat, innovació i investigació en docència universitària. Convocatòria 2020-21

Organització: Institut de Ciències de l'Educació (Vicerectorat de Transformació Digital) de la Universitat d'Alacant/ *Organización: Instituto de Ciencias de la Educación (Vicerrectorado de Transformación Digital) de la Universidad de Alicante*

Edició / *Edición*: Rosana Satorre Cuerda (Coord.), Asunción Menargues Marcillas, Rocío Díez Ros, Neus Pellin Buades

Revisió i maquetació: ICE de la Universitat d'Alacant/ *Revisión y maquetación: ICE de la Universidad de Alicante*

Primera edició / *Primera edición*: desembre 2021/ diciembre 2021

ice@ua.es

ISBN: 978-84-09-34941-8

Producció: Institut de Ciències de l'Educació (ICE) de la Universitat d'Alacant / *Producción: Instituto de Ciencias de la Educación (ICE) de la Universidad de Alicante*

Aquesta publicació s'ha fet seguint les directriu d'accessibilitat UNE-EN 301549:2020 / Esta publicación se ha hecho siguiendo las directrices de accesibilidad UNE-EN 301549:2020.

138.Evaluación cuantitativa de las especificaciones de diseño en problemas de síntesis de sonido mediante aprendizaje automático

Jose J. Valero-Mas; María Alfaro-Contreras; José M. Iñesta; Pedro J. Ponce de León; Francisco J. Castellanos; Jorge Calvo-Zaragoza

jjvalero@dlsi.ua.es (Jose J. Valero-Mas), malfaro@dlsi.ua.es (María Alfaro-Contreras); inesta@dlsi.ua.es (José M. Iñesta); pierre@dlsi.ua.es (Pedro J. Ponce de León), fcastellanos@dlsi.ua.es (Francisco J. Castellanos), jcalvo@dlsi.ua.es (Jorge Calvo-Zaragoza)

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante

Carretera San Vicente del Raspeig s/n, Alicante, 03690, Spain

Resumen

La evaluación de las tareas en problemas de síntesis de sonido constituye un reto dada la naturaleza acústica de la salida, forzando generalmente a un proceso de corrección manual con las desventajas de subjetividad y cansancio que ello conlleva. En este trabajo proponemos un modelo que sea capaz de evaluar este tipo de problemas de manera automática y así apoyar al docente en esta tarea con la finalidad de que pueda destinar ese tiempo a otro tipo de tareas docentes con más impacto en el proceso de enseñanza-aprendizaje del alumno. La propuesta se basa en modelos de aprendizaje neuronal profundo, concretamente las llamadas Redes Siamesas, los cuales permiten medir la similitud entre dos elementos de entrada. En nuestro caso, en base a un grupo de datos etiquetados como correctos e incorrectos, el modelo nos permite estimar a qué tipo de elemento se asemejan más las diferentes entregas de los alumnos. Los resultados obtenidos con una serie de escenarios simulados en base a entregas de estudiantes de la asignatura Síntesis Digital de Sonido del Grado en Ingeniería en Sonido e Imagen en Telecomunicación validan la propuesta realizada en todos los casos propuestos.

Palabras clave: Evaluación cuantitativa; Aprendizaje automático; Síntesis de sonido; Redes neuronales

1. Introducción

La Síntesis Digital de Sonido es la disciplina dentro del campo conocido como Extracción y Recuperación de Información Musical (del inglés *Music Information Retrieval*) que investiga y diseña métodos y algoritmos capaces de crear, por medios puramente computacionales, sonidos tanto existentes en la naturaleza como aquellos que sólo existen en la mente de las personas (Serra et al., 2013). En nuestro día a día encontramos una gran cantidad de casos en los que se aplica síntesis de sonido, como pueden ser películas, series de televisión, videojuegos y realidad virtual, entre otros (Merer et al., 2013).

Dado que el producto que se busca obtener con este tipo de medios es sonido y/o música, es lógico pensar que los principales usuarios serían artistas sin ningún tipo de restricciones a la hora de crear más allá de las puramente estéticas. Sin embargo, en ciertas ocasiones esta creación no es de una naturaleza tan libre y su diseño se ha de regir por una serie de especificaciones que, normalmente, llevan a cabo ingenieros. Es por ello que no es extraño encontrar planes de estudios universitarios que cuentan con esta materia, como pueden ser el Grado en Ingeniería en Sonido e Imagen en Telecomunicación (Universidad de Alicante) o el Grado en Ingeniería en Sistemas Audiovisuales (Universitat Pompeu Fabra), entre otros. Es en este contexto en el que se centra el presente trabajo.

Como cualquier otra disciplina dentro de la rama de la ingeniería, la Síntesis Digital de Sonido se debe regir por unas especificaciones claras que posteriormente puedan ser cuantitativamente medibles y evaluables para poder comprobar si el diseño que se ha llevado a cabo ha seguido estos criterios o no, al menos en un contexto educativo. Sin embargo, al contrario de otras materias como puede ser la Programación dentro del Grado en Ingeniería Informática en que es evaluar la eficiencia y eficacia de un código de programación es relativamente sencillo y, hasta cierto punto, autónomo, o del Cálculo de Estructuras en el Grado en Arquitectura cuya evaluación es la misma que la de un problema de física, para la Síntesis de Sonido no resulta sencillo proponer

una evaluación cuantitativa debido a la naturaleza acústica de la salida. En ese sentido el proceso de evaluación de este tipo de asignaturas en los diferentes centros en los que se estudia tiene que ser de tipo cualitativo, lo cual conlleva tendencia a errores por factor humano además de baja escalabilidad de la estrategia.

Es en este contexto en el que el presente trabajo busca realizar una primera aproximación, desde un punto de vista educativo, hacia la cuestión de la evaluación cuantitativa de las especificaciones de diseño en problemas de Síntesis Digital de Sonido para estudiantes de la rama de ingeniería.

1.1 Problema o cuestión específica del objeto de estudio

Como se ha comentado, el contexto de trabajo es la evaluación de los problemas de Síntesis Digital de Sonido en un contexto de ingeniería.

De la misma manera que ocurre en otros problemas de estas ramas, normalmente los diseños a realizar están sujetos a una serie de especificaciones de cuyo cumplimiento resulta el éxito de la tarea en sí. Sin embargo, como se ha mencionado ya, la naturaleza acústica de este tipo de tareas dificulta mucho su evaluación de manera cuantitativa.

La cuestión concreta que busca atacar este trabajo es la de evaluar si, por medios basados en aprendizaje automático, es posible establecer si el sonido resultante de un problema de síntesis llevado a cabo por un alumno cumple las especificaciones impuestas por el profesor sin necesidad de otras ayudas adicionales y de manera autónoma. Esto no sólo permitiría una evaluación totalmente objetiva sino también escalable a cualquier número de estudiantes.

1.2 Revisión de la literatura

Dada la importancia de la componente artística y creativa en las tareas de síntesis de sonido, su evaluación ha sido siempre un tema controvertido por la

dificultad de encontrar una metodología clara que pudiera responder a esta necesidad.

El trabajo de Jaffe (1995) supone una primera aproximación a este problema. En él se proponen diez criterios diferentes para la evaluación de sonidos sintetizados (cinco referidos a la evaluación de los parámetros de control, tres al propio coste computacional de la tarea de síntesis y dos destinados a evaluar las cualidades del sonido creado en sí). Tolonen, Välimäki y Karjalainen (1998) expandieron este estudio incorporando una serie adicional de sistemas de síntesis de sonido.

En un plano centrado en la evaluación perceptual de la calidad de la síntesis de sonido, Moffat y Reiss (2018) realizaron una profunda revisión de la literatura asociada además de proponer una metodología para evaluar ese aspecto en la síntesis de efectos sonoros. A pesar de la validez científica de esta propuesta, esta evaluación no es aplicable para nuestro concreto al basarse en una validación de tipo perceptual y no basada en especificaciones.

El antecedente más directo hacia la evaluación automática de este tipo de problemas, aunque aunque no es posible hacer referencia a una publicación formal a la misma al no haber sido publicada en ningún medio científico, la encontramos en una primera aproximación realizada en la asignatura Síntesis Digital de Sonido del Grado en Ingeniería en Sonido e Imagen en Telecomunicación de la Universidad de Alicante. En la parte práctica de esta asignatura, y en base a las especificaciones de diseño del profesor, los alumnos llevan a cabo sus respectivos diseños de síntesis sonido mediante el lenguaje de programación Csound (Vercoe, 1986). Este lenguaje, además de proporcionar el fichero sonoro resultante del proceso de síntesis, también es capaz de devolver una serie de parámetros resultantes del propio proceso de síntesis, como son la duración del sonido, la amplitud máxima de la señal sonora o la cantidad de muestras sonoras saturadas, entre otros datos.

En base a ello, los profesores de la asignatura en los cursos 2015 – 2016 y 2016 – 2017 (José J. Valero-Mas, Pedro J. Ponce de León y José M. Iñesta) desarrollaron un programa que, de manera automática, se encargaba sintetizar los diferentes sonidos resultantes de las implementaciones de los alumnos y de recoger los parámetros anteriormente citados. La divergencia entre los valores

de los parámetros por los alumnos y los obtenidos en la implementación canónica realizada por los profesores se utilizó como estimación de la nota del estudiante. A pesar de que los resultados entre la corrección basada en el programa anteriormente comentada y la corrección manual guardaban una alta correlación, la corrección se basaba en la extracción de parámetros muy genéricos de la señal acústica, siendo necesario un siguiente avance que permita evaluar la consecución de las especificaciones de diseño.

Actualmente, el avance de los modelos basados en Aprendizaje Profundo (del inglés, *Deep Learning*) han supuesto un gran avance en las diferentes disciplinas derivadas y que hacen uso del Aprendizaje Automático (LeCun, Bengio y Hinton, 2015). El campo de la Extracción y Recuperación de Información Musical no ha sido ajeno a estos avances, representando este campo una de las principales herramientas para la resolución de problemas de índole musical, tales como análisis tonal, composición algorítmica o efectos de audio, entre otros.

De la misma manera que en los casos mencionados, el campo del Aprendizaje Profundo también está influyendo el área de la síntesis de sonido. A diferencia de los sistemas clásicos basados en algoritmos y métodos fijos, las nuevas propuestas de síntesis basadas en la mencionada rama permiten obtener sonidos con resultados increíblemente realistas y fidedignos. Algunos ejemplos los encontramos en los trabajos de Engel y otros (2018) en el que se utilizan Redes Generativas Adversarias (del inglés, *Adversarial Neural Networks*) para generar sonidos con arquitecturas prácticamente no supervisadas o el trabajo de Liu y Manocha (2020) que realiza una revisión de la literatura sobre sistemas de síntesis basados en Aprendizaje Automático Profundo.

Dentro de ese campo también han aparecido una serie de modelos neuronales que nos permiten comparar señales (normalmente imágenes) y obtener un valor de similitud como resultado de esa comparación. Este tipo de arquitecturas, llamadas Redes Siamesas (del inglés, *Siamese Neural Networks*) (Bromley y otros, 1993), constituyen la base técnica de esta red docente y por ello serán introducidas en las siguientes secciones.

1.3 Objetivos del estudio

A continuación, listamos los objetivos concretos que se busca conseguir con este trabajo:

- Creación de un corpus de datos con los resultados de programas de síntesis de sonido de diferentes alumnos.
- Creación de un corpus de datos con las versiones canónicas de los profesores de la asignatura correspondientes a los ejercicios asignados a los alumnos.
- Implementación de un modelo neuronal, que será descrito en el siguiente apartado, capaz de evaluar la similitud entre dos señales.
- Evaluación del modelo neuronal como método para medir la consecución de las especificaciones de diseño.

2. Método

2.1. Descripción del contexto y de los participantes

Dentro de los estudios de la Universidad de Alicante existen tres asignaturas que tratan esta temática: Síntesis Digital de Sonido (20029), del Grado en Ingeniería en Telecomunicación en Sonido e Imagen, cuyo carácter es optativo y se oferta en el tercer curso; Sonido y Música por Computador (21028) del tercer curso del Grado en Ingeniería Multimedia y de carácter obligatorio y Técnicas de Diseño Sonoro (21039) del mismo grado pero de cuarto curso y carácter optativo. En todas ellas gran parte de la evaluación consiste en la síntesis de sonidos por medios puramente computacionales y, aunque cada una de las asignaturas se centra en aspectos diferentes de la disciplina de la síntesis de sonido, existen ciertas partes del temario que son comunes a las tres, además de utilizar todas ellas el lenguaje de programación Csound para el desarrollo de la parte práctica.

En nuestro caso nos vamos a centrar en la asignatura Síntesis Digital de Sonido (20029) que, como se ha comentado, pertenece al Grado en Ingeniería

en Telecomunicación en Sonido e Imagen. Esta asignatura es optativa, de 6 créditos ECTS de duración, se cursa en el segundo cuatrimestre del tercer año de la titulación y es heredera de la asignatura del mismo nombre del extinto plan de Ingeniería Técnica en Telecomunicación: especialidad en Sonido e Imagen de la Universidad de Alicante, que también tenía un carácter optativo y una carga lectiva de 7,5 créditos ECTS.

Cabe destacar que los estudiantes de esta materia comparten un perfil bastante similar que puede ser relevante a la hora de extraer conclusiones sobre este estudio. Dado que es una asignatura optativa cuyo foco principal es la música desde un punto de vista de ingeniería, un alto porcentaje de los alumnos han recibido de alguna manera formación musical (conservatorios, escuelas de música o incluso de manera autodidacta) y tocan algún instrumento musical. Además, dado que todos son estudiantes de ingeniería, se presupone una serie de conocimientos técnicos adquiridos durante la titulación, especialmente referidos a temas de programación. Por último, el hecho de que la asignatura sea optativa, ha permitido observar, a lo largo de los años, que suele haber dos grupos de notas claramente diferenciados: notas relativamente altas, que es la gente con mucho interés musical y tecnológico que son capaces de exprimir la asignatura al máximo exponente, y otro grupo con notas bajas (no suspendidos, pero sí aprobados con poco margen), que suele corresponder a gente que cursa la materia por el hecho de conseguir los créditos necesarios para cumplir expediente y finalizar la titulación.

Finalmente, para el estudio propuesto en este trabajo, hemos seleccionado los alumnos del presente curso 2020 – 2021 de la asignatura Síntesis Digital de Sonido. En total constituyen una población de 28 estudiantes, 21 hombres y 7 mujeres, y 12 prácticas por estudiante en las que cada una busca desarrollar un concepto concreto de síntesis de sonido explicado en las sesiones de teoría.

2.2. Instrumento utilizado para realizar la investigación

Como se ha comentado, esta investigación se centra en una asignatura de una titulación universitaria de la rama de la ingeniería. En ella, las prácticas

buscan crear sonidos por medios computacionales mediante un código de programación, concretamente el lenguaje Csound. En este sentido el instrumento utilizado para realizar la investigación ha sido las diferentes entregas que cada alumno ha realizado para la consecución de la asignatura. A modo de ejemplo, la Figura 1 muestra un ejemplo de código de programación en Csound.

```
ksmps = 180  ; las variables k- se recalculan cada 180 muestras
sr = 44100             ; frecuencia de muestreo en Hz
nchnls = 2             ; sonido en mono
0dbfs = 1      ; Valor de amplitud normalizado entre -1 y 1

instr principio

      prints "Esperamos 1 s\n"
      giFrec = cpspch(9.07)

endin

instr creciente

      print giFrec
```

Figura 1. Ejemplo de código del lenguaje de programación de Csound sobre una de las prácticas de la asignatura Síntesis Digital de Sonido del Grado en Ingeniería en Sonido e Imagen en Telecomunicación.

Por otro lado se ha utilizado la plataforma Moodle para gestionar todas la realización y entrega de prácticas. Esta plataforma permitía especificar los enunciados para cada uno de los ejercicios de los alumnos y, una vez realizados en el ordenador de manera local, realizar una entrega con todos los ejercicios juntos. La Figura 2 muestra un ejemplo de tarea en esta plataforma.

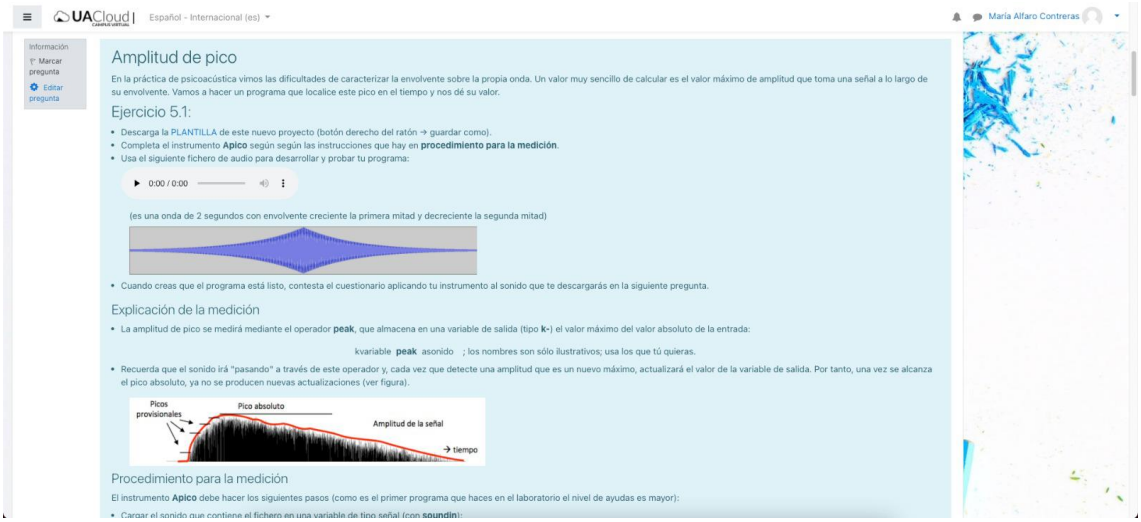

Figura 2. Ejemplo de tarea planteada al alumno en la plataforma Moodle.

2.2.1 Redes Neuronales Siamesas

La base técnica de este estudio son las Redes Neuronales Siamesas (Bromley y otros, 1993) anteriormente introducidas. Como se ha comentado, se trata de un tipo de modelo que, dado un par de señales de entrada, nos permite obtener un valor que representa el grado de similitud entre las mismas.

Para realizar esta tarea, este tipo de arquitecturas neuronales siguen este proceso: dado un par de datos de entrada I_a e I_b, la red realiza un cambio de representación a un nuevo espacio vectorial en el que estos datos de entrada se representan como los vectores X_a y X_b; en este nuevo espacio vectorial es sencillo definir una medida de distancia que hace las veces de medida de similitud entre los elementos, siendo típicamente utilizada la distancia euclídea; valores de distancia bajos representarán un alto grado de similitud entre las señales de entrada mientras que distancias altas representan elementos de entrada muy diferentes. La Figura 3 representa esquemáticamente esta arquitectura.

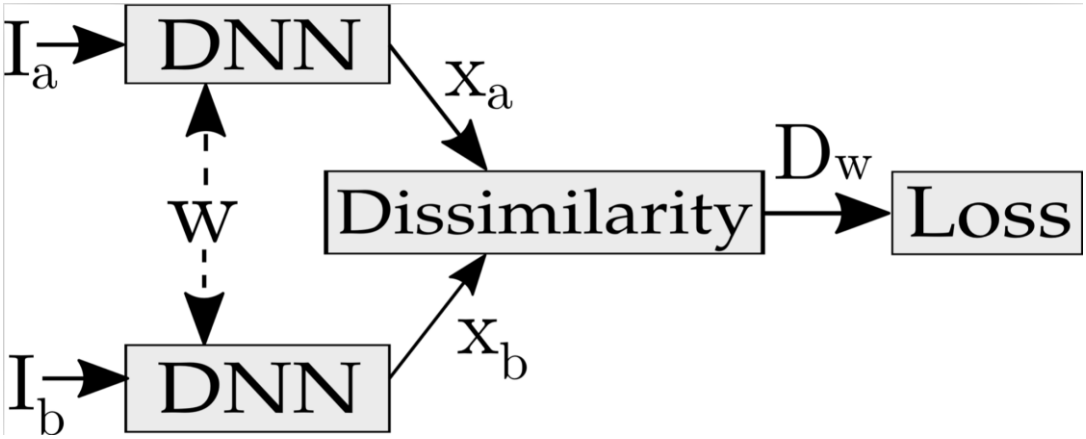

Figura 3. Diagrama de una Red Neuronal Siamesa. Los elementos de entrada I_a e I_b son transferidos a los vectores X_a y X_b, respectivamente, por medio de unas Redes Neuronales Profundas (del inglés *Deep Neural Networks*, DNN). Con una función de disimilitud (*Dissimilarity* en el esquema), típicamente la distancia euclídea, obtenemos una valor D_W de cuán similares son las dos

muestras de entrada. Este valor, además, se utiliza en el entrenamiento del modelo como valor de pérdida o *loss*.

Como se ha comentado, esta investigación busca comprobar si, mediante este tipo de arquitecturas, podemos estimar la similitud entre el ejercicio realizado por un estudiante y la solución canónica del profesor. En este contexto, cuanto más se alejen entre sí ambas señales (medida de disimilitud o distancia altas), la premisa es que menos se parecerán ambos sonidos. Entonces, la idea será ver si este valor está correlacionado con la corrección manual del profesor y así poder obtener un método de evaluación de las señales sin tener que escucharlas.

Cabe destacar que este tipo de arquitecturas están ideadas para procesamiento de imágenes. En ese sentido, para poder trabajar con ficheros sonoros, es necesario realizar una transformación inicial de los datos para obtener una representación adecuada al problema. Inspirados en el campo de la Extracción y Recuperación de Información Musical, obtenemos el módulo del espectrograma de la señal de audio y lo guardamos como imagen. De esta manera podemos obtener una imagen que representa la señal de audio además de ser un tipo de codificación que facilita el aprendizaje de la red al mostrar directamente el contenido frecuencial de la misma.

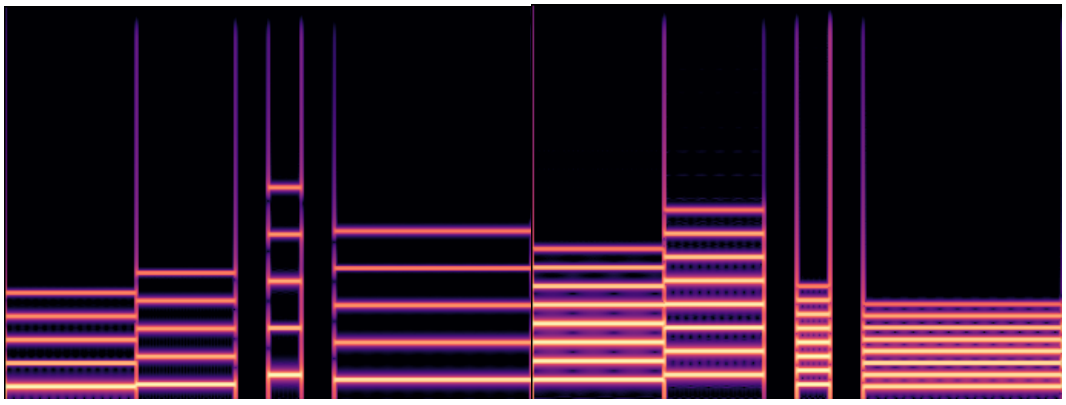

Figura 4. Ejemplo de representación en formato espectrograma de dos señales de sendos ejercicios escogidos de manera aleatoria de los diferentes planteados en la asignatura.

2.3. Procedimiento

El procedimiento concreto ideado para la investigación educativa ha conllevado las siguientes etapas:

- Para cada práctica de la asignatura que busca reforzar un concepto impartido en la asignatura, los diferentes profesores definieron una serie de problemas de síntesis de sonido con unas especificaciones claras y sin ambigüedades. Con esta definición como referencia, se generaron los programas en código Csound que permitían sintetizar los sonidos considerados como correctos, así como códigos en cuya síntesis permite obtener sonidos que considerarán como totalmente incorrectos. Además, también se han definido y generado casos en los que las especificaciones se cumplen únicamente de manera parcial y así definir casos parcialmente correctos.
- Con esta serie de muestras se ha entrenado un sistema de aprendizaje automático como el descrito en la Sección 1.2 capaz de medir la similitud entre señales de manera automática.
- Durante el transcurso de la asignatura se han planteado las diferentes tareas de síntesis de sonido a los alumnos para así obtener las muestras de implementaciones reales de códigos realizados por estudiantes.
- Las diferentes entregas de los alumnos han sido evaluadas a mano por los docentes de la asignatura. Cada una de las prácticas ha sido evaluada por los diferentes docentes que la imparten pero, a fin de disminuir la posible dispersión en la nota asignada, ha sido validada por el resto de docentes de la misma.
- Por medio de la proyección de los sonidos de los alumnos al espacio que describe el algoritmo de aprendizaje se ha buscado ajustar una función que permita relacionar la nota obtenida con la distancia entre las muestras de referencia de los profesores.

3. Resultados

El experimento que se ha llevado a cabo para verificar la utilidad de la propuesta ha consistido en escoger un ejercicio concreto de los planteados en la asignatura y crear una serie de variaciones del mismo. Por simplicidad hemos seleccionado un ejercicio de las primeras prácticas de la asignatura, que son conceptualmente sencillas y con una rúbrica sencilla de corrección. La partitura de este ejercicio se muestra en la Figura 5.

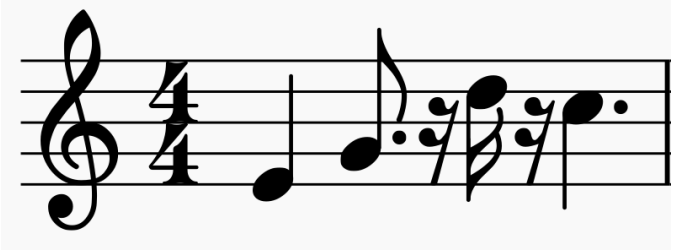

Figura 5. Ejercicio a sintetizar por medio de código CSound utilizado para la evaluación de la propuesta docente.

En base a este enunciado se han generado cinco sonidos base para experimentar si el modelo neuronal puede detectar las variaciones entre ellos:

- **Sonido correcto:** La resolución correcta del ejercicio planteado.
- **Incorrecto #1:** Simulación de resolución incorrecta de la práctica en la que la progresión creada está transpuesta una segunda hacia arriba.
- **Incorrecto #2:** Simulación de resolución incorrecta de la práctica en la que la progresión creada es la misma que la correcta pero en orden inverso.
- **Incorrecto #3:** Simulación de resolución incorrecta de la práctica en la que la progresión creada está transpuesta una octava hacia abajo.
- **Incorrecto #4:** Simulación de resolución incorrecta de la práctica en la que se ha sintetizado únicamente la primera nota de la secuencia.

Además, para incluir mayor variabilidad en los experimentos, estos cinco casos se han sintetizado con cuatro timbres diferentes cada uno: onda sinusoidal, onda cuadrada, diente de sierra y pulso. Estos timbres corresponden a formas de onda típicas con las que se instruye en esta asignatura, por lo que es interesante estudiar el comportamiento del sistema en este contexto.

Para los experimentos se ha considerado que cada una de las muestras anteriormente introducidas puede formar parte o no de la fase de entrenamiento del modelo de similitud. Dado que ambas situaciones son perfectamente factibles (dos alumnos que realicen un ejercicio igual sin ningún tipo de copia), las evaluaremos en nuestros experimentos. Los resultados en términos de disimilitud o distancia obtenidos bajos estas consideraciones para diferentes combinaciones de tipos de sonidos se muestran en la Tabla 1.

Tabla 1. Resultados, en términos de disimilitud o distancia, obtenidos por el modelo neuronal al comparar dos muestras de datos. Los términos *visto* y *no visto* representan los casos en los que la muestra en cuestión ha formado parte de la fase de entrenamiento del modelo o no, respectivamente.

Caso	Muestra 1	Muestra 2	Disimilitud
1	Correcto (no visto)	Incorrecto (no visto)	0,686
2	Correcto (visto)	Incorrecto (no visto)	0,612
3	Correcto (visto)	Correcto (no visto)	0,294
4	Incorrecto (visto)	Incorrecto (no visto)	0,102
5	Correcto (visto)	Incorrecto (visto)	0,548
6	Correcto (visto)	Correcto (visto)	0,308

Como se puede observar, en los casos en los que una muestra catalogada como *Correcta* se compara con una *Incorrecta* es grado de disimilitud es mayor

que en aquéllos en los que dos muestras del mismo tipo (ambas de tipo *Correcto* o *Incorrecto*). Este primer resultado valida, hasta cierto punto, la premisa que se buscaba en este trabajo: estudio de la similitud entre muestras de audio mediante métodos neuronales para estimar la realización de especificaciones de diseño. Cabe destacar, además, que una práctica habitual en este tipo de esquemas es establecer un umbral a partir del cual los elementos se consideran diferentes; en este sentido, si establecemos valor $\theta = 0{,}5$ como umbral, podemos ver que los casos 1, 2 y 5 ciertamente serían catalogados como incorrectos mientras que el resto, al no superar ese umbral de disimilitud, estarían catalogados como correctos.

Una vez introducidos los resultados de manera general, a continuación realizaremos un análisis pormenorizado de los mismos ya que cada uno modela una situación concreta que se puede dar en un aula:

- **Caso 1:** El primero de los escenarios muestra un caso en que un sonido *Correcto* se compara a uno *Incorrecto*, siendo ninguno de ellos visto en la fase de entrenamiento. Nótese que esta situación representa un caso que se podría dar en la práctica ya que una tarea incorrecta no vista en la fase de entrenamiento se está comparando con una tarea correcta tampoco utilizada para entrenar el sistema. Como se ha comentado, la distancia obtenida entre las muestras permite ver que la solución planteada por el alumno no es correcta.
- **Caso 2:** Esta segunda situación es análoga al escenario anterior pero plantea una situación más realista ya que la muestra *Correcta*, que es la que realiza el profesor, está utilizada en el entrenamiento. Como en el caso anterior, se demuestra que la propuesta es capaz de detectar que la tarea es incorrecta dada la distancia obtenida entre las muestras.
- **Caso 3:** Este escenario constituye también un caso realista ya que una respuesta *Correcta* no vista en entrenamiento se enfrenta a la solución *Correcta* del profesor que, además, ha sido utilizada para entrenar el sistema. De nuevo, el bajo valor de distancia implica un alto grado de similitud, lo que valida la técnica propuesta.
- **Caso 4:** En este planteamiento se enfrentan dos muestras *Incorrectas*, siendo una utilizada en entrenamiento y otra no. Esta situación también

constituye un caso que se puede dar en la realidad ya que el profesor puede generar muestras incorrectas en función de su experiencia y de los errores vistos en otros cursos. Por otro lado, la muestra *Incorrecta* no vista en entrenamiento representa la entrega de un alumno. Como se puede observar el bajo grado de disimilitud entre las muestras dice que ambas son equivalentes, es decir, incorrectas.

- **Caso 5:** En esta situación mostramos un caso realizado únicamente para la completitud de resultados ya que representa un caso que no se puede dar en la realidad. Concretamente, tenemos una muestra *Correcta* y otra *Incorrecta*, ambas vistas durante la fase de entrenamiento. Como se puede observar, la distancia resultante establece que ambos casos son diferentes, que era el resultado esperado.
- **Caso 6:** Como en el escenario anterior, este planteamiento también representa un caso irreal realizado para evaluar todas las posibles combinaciones. Como se puede ver, ambas muestras son *Correctas* y han sido vistas durante el entrenamiento, lo que devuelve un grado de similitud alto.

En base a estos resultados podemos ver que la metodología propuesta responde perfectamente a la necesidad planteada de estimar si dos sonidos resultantes de procesos de síntesis diferentes representan o no la misma especificación. Esta validación de la propuesta se traduce en una mejora de la calidad docente ya que, como se ha introducido anteriormente en la memoria, la correción por medios automáticos garantiza un proceso objetivo, rápido y sin fallos humanos. En ese sentido el docente puede dedicar su tiempo a generar tareas que constituyan retos reales e interesantes a los estudiantes en lugar de deber dedicar parte de su tiempo a esa corrección. Más aún, en muchas ocasiones los diseños de prácticas se ven limitados por lo tedioso que puede ser la corrección de esas tareas; automatizando este proceso, como se sugiere en este trabajo, se podría conseguir que el docente ya no tuviera esas conductas de evitamiento.

Sin embargo, estos resultados son todavía preliminares. Por ejemplo, la mayor deficiencia a solventar aquí es el ser capaz de que el modelo pueda dar

razones sobre la decisión tomada, la llamada *interpretabilidad* del modelo (Burkart y Huber, 2021). Como se ha observado, la propuesta es capaz de esclarecer si un sonido sintetizado es correcto en base a compararlo con otros que ya sabemos si lo son; el problema es que, sea la decisión que sea, el método no es capaz de explicar hasta qué punto se han cumplido las especificaciones.

4. Conclusiones

La Síntesis Digital de Sonido es una disciplina dentro del campo del Procesado de Señales que busca crear, mediante medios puramente computacionales, sonidos. A pesar de su inherente componente artística, esta materia es estudiada en diferentes titulaciones relacionadas con la ingeniería debido a que también se trata de un proceso que se puede realizar en base a especificaciones. Sin embargo la evaluación de este tipo de tareas, al menos en un ámbito docente, resulta ser una tarea ardua y tediosa por lo particular de su salida, además de ser propensa a errores y, hasta cierto punto, subjetiva.

Este trabajo propone una metodología para solventar ese tipo de problemas en la evaluación de tareas de este campo en ámbitos educativos con el objetivo de ayudar al docente en la tarea de corrección y así poder destinar ese tiempo a otras tareas docentes que mejoren la experiencia del alumno. Para ello hemos propuesto un sistema basado en modelos de redes neuronales capaz de medir la similitud entre muestras de audio que permite estimar si un sonido es correcto o no en base a una colección de sonidos de referencia.

Esta propuesta ha sido evaluada con una serie de experimentos que emulan posibles escenarios de ejercicios entregados y elementos de referencia. Tomando la precaución de que esta experiencia constituye un ensayo a nivel de laboratorio, los resultados preliminares obtenidos validan la propuesta realizada en todos los casos propuestos.

Como trabajo futuro proponemos seguir explorando este modelo, u otros alternativos, para poder obtener una mejor justificación de los resultados que se obtienen. Además, también es de nuestro interés ampliar la experimentación

realizada de cara a validar la propuesta en entornos menos controlados y así descubrir otras limitaciones del modelo propuesto.

5. Tareas desarrolladas en la red

El desarrollo de las tareas de recopilación de datos, discusión de ideas, experimentación, validación y análisis de resultados se ha llevado a cabo por los seis integrantes de esta red docente. En este caso, todos pertenecen al Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante. Cuatro de ellos son o han sido docentes de la asignatura en cuestión, mientras que los otros dos integrantes son expertos en Aprendizaje Profundo, que constituye la base técnica de esta red docente. A continuación se detallan los nombres de estos componentes y las tareas desarrolladas por cada uno.

Participante de la red	Tareas que desarrolla
Jose J. Valero-Mas	Coordinación, implementación e investigación. Aportación de ideas base. Redacción de la memoria.
María Alfaro-Contreras	Recogida de datos, evaluación de los mismos y aportación de ideas base.
José M. Iñesta	Recogida de datos, evaluación de los mismos y aportación de ideas base.
Pedro J. Ponce de León	Aportación de ideas base.
Francisco J. Castellanos	Aportación de ideas base.
Jorge Calvo-Zaragoza	Aportación de ideas base.

6. Referencias bibliográficas

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature verification using a" siamese" time delay neural network. *Advances in Neural Information Processing Systems*, 6, 737-744.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245-317.

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2018, September). GANSynth: Adversarial Neural Audio Synthesis. In *International Conference on Learning Representations (ICLR)*.

Jaffe, D. A. (1995). Ten criteria for evaluating synthesis techniques. *Computer Music Journal*, 19(1), 76-87.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Liu, S., & Manocha, D. (2020). Sound Synthesis, Propagation, and Rendering: A Survey. *arXiv preprint arXiv:2011.05538.*

Merer, A., Aramaki, M., Ystad, S., & Kronland-Martinet, R. (2013). Perceptual characterization of motion evoked by sounds for synthesis control purposes. *ACM Transactions on Applied Perception (TAP),* 10(1), 1-24.

Moffat, D., & Reiss, J. D. (2018). Perceptual evaluation of synthesized sound effects. *ACM Transactions on Applied Perception (TAP)*, 15(2), 1-19.

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., ... & Widmer, G. (2013). *Roadmap for music information research*.

Tolonen, T., Välimäki, V., & Karjalainen, M. (1998). *Evaluation of modern sound synthesis methods*. Helsinki University of Technology.

Vercoe, B. (1986). Csound. *The CSound Manual Version*, 3.