UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS

FACULTAD DE INGENIERÍA

PROGRAMA ACADÉMICO DE INGENIERÍA ELECTRÓNICA

Aplicación de redes neuronales convolucionales para la emulación del modelo
psicoacústico MPEG-1, capa I, para la codificación de señales de audio

TESIS

Para optar el título profesional de Ingeniero Electrónico

AUTOR(ES)

Sanchez Huapaya, Alonso Sebastián (0000-0002-4449-786X)

Serpa Pinillos, Sergio André (0000-0001-5794-4756)

ASESOR

Kemper Vásquez, Guillermo Leopoldo (0000-0002-7800-7769)

Lima, 26 de agosto de 2020

AGRADECIMIENTOS

A mis padres, Manuel y Luisa, por haberme dado la oportunidad de formarme en esta prestigiosa universidad y haber sido mi apoyo durante todo este tiempo.

De manera especial a mi asesor Dr. Guillermo Kemper, por haberme guiado, no sólo en la elaboración de este trabajo de titulación, sino a lo largo de mi carrera universitaria y haberme brindado el apoyo para desarrollarme profesionalmente y seguir cultivando mis valores.

A la Universidad Peruana de Ciencias Aplicadas, por haberme brindado tantas oportunidades y enriquecerme en conocimiento.

Sergio

A Dios, por la vida y oportunidades presentadas.

A mi familia, especialmente a mi madre Liliana, por el apoyo, paciencia y comprensión.

A mi enamorada Bárbara, por acompañarme tan de cerca en la travesía de mi titulación.

A mi asesor Dr. Guillermo Kemper, por el apoyo y orientación en las diversas facetas de la elaboración de la tesis.

Alonso

RESUMEN

El presente trabajo propone 4 alternativas de codificadores inspirados en el codificador MPEG-1, capa I, descrito en el estándar ISO/IEC 11172-3. El problema que se intenta resolver es el de requerir definir un modelo psicoacústico explícitamente para lograr codificar audio, reemplazándolo por redes neuronales. Todas las alternativas de codificador están basadas en redes neuronales convolucionales multiescala (MCNN) que emulan el modelo psicoacústico 1 del codificador mencionado. Las redes tienen 32 entradas que corresponden a las 32 subbandas del nivel de presión sonora (SPL – sound pressure level), y una única salida que corresponde a una de las 32 subbandas de o bien la relación señal a máscara (SMR) o bien el vector de asignación de bits. Es decir, un codificador está compuesto de un conjunto de 32 redes neuronales. La validación empleó los 10 primeros segundos de 15 canciones elegidas aleatoriamente de 10 géneros musicales distintos. Se comparó la calidad de las señales de audio generadas por cada codificador contra la de MPEG-1, capa I, mediante la métrica de ODG. El codificador cuya entrada es el SPL y cuya salida es la SMR, planteado por Guillermo Kemper, obtuvo los mejores resultados al realizar la comparación para 96 kbps y 192 kbps. El codificador denominado “SBU1” obtuvo los mejores resultados para 128 kbps.

Palabras clave: codificador; modelo psicoacústico; redes neuronales convolucionales multiescala; subbandas; relación señal a máscara; señales de audio; nivel de presión sonora.

Convolutional neural networks applied to the emulation of the psychoacoustic model for
MPEG-1, Layer I audio signal encoders

ABSTRACT

The present work proposes 4 encoder alternatives, inspired in the MPEG-1, layer I encoder described in the ISO/IEC 11172-3 standard. The problem addressed here is the requirement of explicitly defining a psychoacoustic model to code audio, instead replacing it by neural networks. All the proposals are based on multiscale convolutional neural networks (MCNN) that emulate the psychoacoustic model 1 of the referred encoder. The networks have 32 inputs that map the 32 subbands of the sound pressure level (SPL), and a single output that corresponds to each of the 32 subbands of either the signal-to-mask ratio (SMR) or the bit allocation vector. Thus, an encoder is composed of a set of 32 neural networks. The validation process took the first 10 seconds of 15 randomly chosen songs of 10 different musical genres. The audio signal quality of the proposed encoders was compared to that of the MPEG-1, layer I encoder, using the ODG metric. The encoder whose input is the SPL and whose output is the SMR, proposed by Guillermo Kemper, yielded the best results for 96 kbps and 192 kbps. The encoder named “SBU1” had the best results for 128 kbps.

Keywords: encoder; psychoacoustic model; multiscale convolutional neural networks; subbands; signal-to-mask ratio; audio signal; sound pressure level.