Procesamiento de Audio con Redes Neuronales

Dr Diego Fernández Slezak (Profesor Adjunto, Depto de Computación, FCEN-UBA) con la colaboración del Dr. Pablo Riera (JTP)

Programa:

Esta materia busca introducir temáticas del procesamiento de habla, desde el punto de vista del procesamiento de señales clásico y también del de las redes neuronales. Se busca que los participantes logren llevarse los conceptos básicos del procesamiento de señales clásico (filtros, análisis de Fourier) para luego pasar a métodos actuales basados en redes neuronales (autoencoders, self-supervised learning) teniendo en cuenta el uso de estas herramientas para tareas de análisis de audio y tareas de reconocimiento del habla.
La materia busca una mezcla balanceada de teoría y ejercitación práctica.

Temario:

Elementos de acústica y procesamiento digital de señales. Elementos de fonología y producción de habla. Análisis de Fourier y filtros. Elementos de estadística y aprendizaje automático. Elementos de procesamiento del lenguaje natural. Reconocimiento del habla. Modelos ocultos de Markov (HMM) y redes neuronales profundas. Aprendizaje de representaciones. Temas avanzados: reconocimiento de eventos sonoros y géneros musicales; extracción de información del hablante; detección de emociones;

- Programa (tentativo, pueden ser más o menos clases, sumado a las clases de laboratorio):

Clase 1: Introducción al Procesamiento de Señales

- ¿Qué es el procesamiento de señales? Señales continuas vs. señales discretas. Muestreo y cuantización.
- Convolución y correlación. Auto-correlación y cross-correlación. Filtrado en el dominio del tiempo.

Clase 2: Análisis en el Dominio de la Frecuencia

- Análisis de Fourier: series de Fourier y transformadas de Fourier. Transformada de Fourier Discreta (DFT) y Transformada Rápida de Fourier (FFT).

Clase 3: Análisis en el Dominio de la Frecuencia

- STFT
- Filtrado y convolución en el dominio de la frecuencia.
- Análisis espectral

Clase 4: Procesamiento Estadístico de Señales

- Modelado de señales con distribuciones de probabilidad. Modelos gaussianos y no gaussianos. Detección y estimación en señales ruidosas. Relación señal-ruido (SNR).
- Introducción al filtrado óptimo y el filtro de Wiener.

Clase 5: Introducción al Procesamiento de Audio y Habla

- Características acústicas de las señales de audio y habla.
- Fundamentos de la producción y percepción de sonidos ambientales, musicales y habla.
- Atributos para el reconocimiento de música y habla. Chromagramas, MFCC.

Clase 5: Descubrimiento de unidades
- Clustering. GMMs
- Segmentación de señales
- Identificación de anomalías

Clase 6: Modelado de series y secuencias
- Modelos ARMA
- Modelos Ocultos de Markov (HMMs)

Clase 7: Reconocimiento de Habla Clásico
- GMM-HMMS
- Decodificación con transductores de estados finitos (FSTs)

Clase 8: Introducción al Aprendizaje Profundo
- Arquitectura de redes neuronales. Backpropagation y descenso de gradiente. Funciones de activación y funciones de pérdida.
- Introducción a redes neuronales profundas para señales. Redes Neuronales Convolucionales (CNNs) y Redes Neuronales Recurrentes (RNNs).
- Atención. Transformers
- Modelado de secuencias: seq2seq, Encoder-Decoder

Clase 9: Aprendizaje Profundo en el Procesamiento de Señales
- Autoencoders profundos para el aprendizaje de características.
- Aprendizaje por transferencia para el procesamiento de señales. Aprendizaje auto-supervisado.

Clase 10: Modelos generativos para síntesis de audio
- Compresión de audio
- Supresión de ruido
- Separación de fuentes

Clase 11: Reconocimiento de Habla 2
- DNN-HMM
- Reconocimiento de habla de extremo a extremo con aprendizaje profundo.

Clase 12: Temas extras
- Reconocimiento de eventos sonoros
- Reconocimiento géneros musicales
- Extracción de información del hablante (identidad, edad, género
- Detección de emociones

Bibliografía:

- Smith, J. O. (2007). Introduction to digital filters: with audio applications (Vol. 2). Julius Smith.
- Daniel Jurafsky & James H. Martin, "Speech and Language Processing" (3er edition). Prentice Hall, 2023. (https://web.stanford.edu/~jurafsky/slp3/)
- Keith Johnson, "Acoustic and Auditory Phonetics" (2nd edition). Blackwell, 2003.
- Jacob Benesty, M. Mohan Sondhi & Yiteng Huang (Eds.), "Springer Handbook of Speech Processing". Springer-Verlag, 2008.

- "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016
- Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S. Y., & Sainath, T. (2019). Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing, 13(2), 206-219.
- Mohamed, A., Lee, H. Y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., ... & Watanabe, S. (2022). Self-supervised speech representation learning: A review. IEEE Journal of Selected Topics in Signal Processing.