FICHA DE LA ASIGNATURA: PROCESAMIENTO DE HABLA
Obligatoria de especialidad 4 ECTS

Equipo docente

José Manuel Pardo Muñoz (2 créditos)
Área de conocimiento: Procesamiento de Habla y Aprendizaje automático
Tlf: +34-910672223
Correo: josemanuel.pardom@upm.es
Página personal: https://blogs.upm.es/gthau/jose-manuel-pardo/
Publicaciones:
https://scholar.google.es/citations?user=zzWzTcEAAAAJ&hl=es

Ricardo de Córdoba Herralde (2 créditos)
Área de conocimiento: Procesamiento de Habla y Aprendizaje automático
Tlf: +34- 910672206
Correo: ricardo.cordoba@upm.es
Página personal: https://blogs.upm.es/gthau/ricardo-cordoba/
Publicaciones: http://scholar.google.es/citations?user=HMKjoW8AAAAJ

Objetivos didácticos

El objetivo básico es conocer los fundamentos y principales técnicas para la conversión texto-habla y el reconocimiento de habla, así como las métricas más relevantes para tomar la decisión y evaluar los sistemas de reconocimiento y conversión texto-habla. Estos objetivos se desglosan en los siguientes objetivos específicos:

1. Conocer los fundamentos del procesamiento del habla, con especial énfasis en las distintas tecnologías para extraer las características físicas de la voz y conseguir un procesamiento robusto frente al ruido.
2. Explicar en detalle los fundamentos de los conversores texto-habla, desde el análisis de texto, la síntesis prosódica y la síntesis segmental, comparando las distintas tecnologías y los enfoques recientes basados en redes neuronales.
3. Explicar los fundamentos del reconocimiento de habla, tratando todas las alternativas, comenzando por el dynamic time warping, los modelos ocultos de Markov (HMM), los modelos híbridos HMM neuronales y terminando con los modelos end-to-end.
4. Presentación de los fundamentos del reconocimiento del locutor y del idioma.
5. Presentar metodologías para la evaluación de sistemas de reconocimiento de habla y conversión texto-habla.
6. Presentar las aplicaciones de las tecnologías del habla: sistemas de diálogo, traductores automáticos del habla, tecnologías del habla para personas con necesidades especiales.
7. Presentar herramientas de código abierto disponibles para realizar los sistemas presentados en la asignatura.

Conocimientos y/o destrezas previas recomendadas

- Conocimientos a nivel de usuario avanzado de informática e Internet.

- Estadística y probabilidades
- Fonética acústica
- Procesamiento del Habla.
- Programación en Python.
- Uso de Google Colab.
- Lectura de textos en inglés.
- Autoaprendizaje y planificación.
- Trabajo en equipo.
- Pensamiento crítico.
- Aprender del error.

Posibles asignaturas del Máster directamente relacionadas con ésta

- Fonética Forense: caracterización y producción de la voz.
- Identificación de locutores.
- Enfoques Avanzados en Lingüística Computacional (Redes Neuronales).

Carga de trabajo/estudio prevista por semana para el alumno

7-8 horas de dedicación durante 12 semanas, repartidas entre el estudio del material docente (40%), el análisis de material adicional (20%), la comunicación con el tutor y los compañeros (10%), la asistencia a seminarios presenciales o en línea (30%). La última semana se dedicará a la práctica final (100%).

Descripción general de la asignatura

Esta asignatura es el núcleo de lo que se llaman las Tecnologías del Habla, centrándonos en los sistemas de reconocimiento de habla y de conversión texto-habla. Esta disciplina combina varios campos como el tratamiento de audio mediante procesamiento de señal, el aprendizaje automático, estadística, etc., y tiene una gran variedad de aplicaciones en el mundo real, donde cada vez se accede a más servicios utilizando el habla utilizando sistemas hombre-máquina.

La asignatura está organizada en 4 partes principales:

En la primera se tratan los fundamentos de procesamiento de habla y el procesamiento de señal. Posteriormente se desarrollan todos los módulos que se necesitan en un sistema de conversión texto-habla, desde el análisis de texto, la síntesis prosódica y la síntesis segmental. En la sesión práctica se describirán las herramientas de código abierto para realizar la conversión texto-habla y se planteará una práctica que debe desarrollar el estudiante

En la segunda parte se estudiarán los fundamentos, los métodos y los algoritmos que se usan en la conversión habla-texto incluyendo la programación dinámica, los modelos de Markov y las redes neuronales. Finalmente se describirán las herramientas de código abierto para realizar la conversión habla-texto y se planteará una práctica que debe desarrollar el estudiante.

En la tercera parte se estudiará de forma simplificada los sistemas de reconocimiento de idioma y reconocimiento de locutor y se expondrán distintas aplicaciones de la tecnología del habla como son los sistemas de traducción de habla, los sistemas conversacionales y la aplicación de la tecnología del habla a personas con necesidades especiales.

En la cuarta parte se realizará una práctica para el desarrollo de una tarea que contendrá conceptos, herramientas y algoritmos previamente estudiados.

Cronograma del curso

Semana	Tema	Objetivos didácticos	Profesor
S1	Fundamentos de procesamiento de habla: Muestreo, Transformada Discreta de Fourier. Escala mel, Cepstrum, MFCC, vocoders, cuantificación vectorial	Objetivo 1	José Manuel Pardo-TU
S2	Conversión texto habla : Análisis de texto y síntesis prosódica	Objetivo 2	José Manuel Pardo
S3	Conversión texto habla: síntesis segmental. Concatenación de Unidades y síntesis HMM. **Primer test**	Objetivo 2	José Manuel Pardo-TU
S4	Conversión texto habla : síntesis segmental: Modelos neuronales y evaluación de sistemas TTS	Objetivo 2 y 5	José Manuel Pardo
S5	Herramientas de código abierto para la conversión texto habla y **Enunciado de la primera práctica. Segundo test**	Objetivos 2 y 7	José Manuel Pardo-TU
S6	Reconocimiento de voz: Introducción y Dynamic time warping. Evaluación de los sistemas de reconocimiento de habla.	Objetivos 3 y 5	José Manuel Pardo
S7	Reconocimiento de voz: Teorema de Bayes y modelos de lenguaje. **Enunciado segunda práctica**	Objetivo 3	Ricardo de Córdoba-TU
S8	Reconocimiento de voz: Modelos de Markov	Objetivo 3	Ricardo de Córdoba
S9	Reconocimiento de voz: Adaptación, Modelos Híbridos HMM-Neuronal y modelos end to end. **Tercer test**	Objetivo 3	Ricardo de Córdoba-TU
S10	Reconocimiento de idioma y de locutor	Objetivos 4	Ricardo de Córdoba
S11	Herramientas de código abierto para el reconocimiento de habla. **Enunciado de la tercera práctica**. **Cuarto test**	Objetivos 2 y 7	Ricardo de Córdoba- TU
S12	Preparación de la práctica final. Trabajo individual del alumno.	Objetivo 6	Ricardo de Córdoba
	Práctica final	Todos los objetivos	Todos
	Examen final	Todos los objetivos	Todos
	Examen extraordinario	Todos los objetivos	Todos

Breve descripción de la Metodología(s) de aprendizaje(s) que se prevé utilizar

El curso se impartirá en el campus virtual de la UCM. Se organizará en semanas. Cada semana se hará una propuesta temática de trabajo y se trabajará en dos partes:

1ªparte: presentación de la propuesta temática para que sirva de guía y punto de partida para la reflexión individual. Esta presentación será teórica en general, aunque se irán presentando los aspectos prácticos.

2ª parte: estudio y actividades de consolidación. Se realizarán a lo largo de la semana por parte de los estudiantes y con la guía de los profesores.

El curso concluirá con la participación en una competición entre grupos de estudiantes.

Las tareas a realizar cada semana son:
1.- Presentación del profesor sobre los objetivos y contenidos del módulo. Clase magistral explicando los conceptos fundamentales. Guía para su estudio.
2.- Estudio, por parte del alumno, de los contenidos básicos. Incluirán lecturas (en modo texto, audio y/o vídeo) y/o ejercicios.
3.- Reflexión entre iguales: uso del foro para preguntar y aclarar cuestiones de los contenidos básicos. El profesor hace el seguimiento sin intervenir excepto que sea necesario porque se observan errores o no se resuelven las dudas.
4.- Ampliación opcional de conocimientos con el material complementario.

Se podrán realizar tutorías síncronas en la que el profesor atenderá las dudas que se planteen en el foro.

Enumeración de las actividades de aprendizaje que se prevén utilizar para las sesiones virtuales

Se utilizarán todas las actividades previstas en la memoria más la participación en una competición entre grupos de estudiantes para aplicar los conocimientos adquiridos en la asignatura:

- Visionado de video/audio lecciones y sesiones síncronas o asíncronas.
- Estudio individual del material básico.
- Lectura y análisis de material complementario.
- Resolución de ejercicios prácticos.
- Comunicación virtual con el profesor.
- Foros y comunicación colaborativa.
- Seminarios presenciales.
- Participación en la competición entre grupos de estudiantes.

Enumeración de las actividades de aprendizaje que se prevén utilizar para las sesiones presenciales

- Descripción de conceptos teóricos.
- Resolución de ejercicios teóricos o prácticos.
- Competición entre grupos de estudiantes y presentación de resultados.

Procedimiento de evaluación

La asignatura se evalúa a partir de las actividades siguientes:

- Exámenes de test 20%
- Memoria de las prácticas y práctica final y presentación de resultados: 45 % de la nota final.
- Examen final: 35 % de la nota final.
- "**Es imprescindible aprobar el examen final presencial para aprobar la asignatura**"

Competencias y destrezas que se desarrollarán

Resultados del aprendizaje:

- Saber definir y describir los fundamentos matemáticos de las representaciones de señal y de las redes neuronales.
- Saber definir y describir los fundamentos de la conversión texto-habla, análisis de texto, síntesis prosódica y síntesis segmental.
- Saber evaluar los sistemas de conversión texto-habla y habla-texto.
- Saber definir y aplicar a tareas de procesamiento del lenguaje los principales modelos de redes neuronales para el procesamiento del lenguaje natural hablado

- Saber utilizar soluciones (bibliotecas o repositorios) consolidadas para el procesamiento del lenguaje hablado, la generación de representaciones gráficas, acceso y extracción de datos de la web y el almacenamiento y gestión de versiones de proyectos software
- Saber definir, describir y aplicar los principales modelos y métodos de representación y procesamiento del lenguaje natural hablado en sus niveles fonético-fonológico, léxico, y discursivo.
- Saber seleccionar críticamente el modelo de representación del conocimiento más adecuado a un problema o aplicación de la tecnología del habla.
- Saber aplicar soluciones software para construir aplicaciones de la tecnología del habla

Competencias Específicas:

- CE10. Saber utilizar con suficiente destreza los principales paquetes de programación para resolver tareas de procesamiento del lenguaje natural en formato texto y voz.
- CE11. Conocer el manejo de las herramientas software existentes para el procesamiento de las producciones lingüísticas en diferentes lenguas (ej. segmentadores, analizadores morfológicos, sintácticos, semánticos).
- CE12. Conocer los fundamentos teóricos y de implementación de las aplicaciones existentes de Lingüística Computacional (ej. traducción automática, agentes conversacionales, recuperación de información, extracción de entidades nombradas o generación de resúmenes)

Procedimiento para mostrar el progreso del alumno

- Boletín de calificaciones de la asignatura virtual.

Mecanismos de comunicación docente

- Foro, correo electrónico de la asignatura virtual y videoconferencia.

Mecanismos de tutorización virtual

- Foro, correo electrónico en la asignatura virtual y videoconferencia.

Mecanismos de contacto

- Foro de la asignatura virtual que podrá complementarse cuando sea necesario con sesiones de videoconferencia. Lo atenderán los profesores de la asignatura.
- Correo electrónico institucional de los profesores en caso de que no tenga acceso al campus virtual.
- Además, el alumno dispone de un servicio de ayuda para las incidencias informáticas de la Universidad en https://ssii.ucm.es/estudiante

Mecanismos de contacto para quejas y sugerencias de la asignatura

- El alumno debe presentar su queja, en primer lugar, al profesor, y hacerle cuantas sugerencias considere oportunas sobre la asignatura, tanto por vía de correo electrónico como por un buzón anónimo de “quejas y sugerencias” en la página de Presentación de la asignatura virtual.
- El alumno se podrá dirigir también a la Coordinación del Máster, así como al representante de alumnos en caso de que su queja o sugerencia no sea atendida.
- Además, el máster dispone de un buzón de quejas y sugerencias en su página web atendido por el Coordinador del máster.

Mecanismos para recoger la opinión de los alumnos sobre la asignatura

- Participación en el programa Docentia de la UCM complementado con una encuesta anónima preparada por la Coordinación del Máster.

Requisitos técnicos especiales (no de campus virtual)

- Ordenador con conexión a internet.

Bibliografía de la asignatura

- "Deep Learning for NLP and Speech Recognition". Uday Kamath, John Liu, James Whitaker. 2019. Edit. Springer. ISBN: 978-3-030-14596-5.
- "Automatic Speech Recognition. A Deep Learning Approach". Dong Yu, Li Deng. 2015. Edit. Springer. ISBN: 978-1-4471-5779-3
- "New Era for Robust Speech Recognition. Exploiting Deep Learning". Shinji Watanabe, Marc Delcroix, Florian Metze, John R. Hershey. 2017. ISBN: 978-3-319-64680-0.
- Spoken Language Processing: A Guide to Theory, Algorithm, and System Development Huang, X., Acero, A., Hon, H.W.: 1st edn. Prentice Hall PTR, USA (2001)
- A Survey on Neural Speech Synthesis, Xu Tan, Tao Qin, Frank Soong, Tie-Yan Li, 2021 https://arxiv.org/abs/2106.15561v3
- Statistical parametric speech synthesis, Heiga Zen, Keiichi Tokuda, Alan W. Black, Speech Communication, Volume 51, Issue 11, 2009