



CEU

*Universidad
San Pablo*

Escuela Politécnica Superior

La inteligencia artificial en el análisis de datos: técnicas y aplicaciones en el uso de la voz

Javier Tejedor Noguerales

Profesor titular

Universidad CEU San Pablo

Festividad de San José

Marzo 2024



CEU | Ediciones

La inteligencia artificial en el análisis de datos: técnicas y aplicaciones en el uso de la voz

Javier Tejedor Noguerales

Profesor titular

Universidad CEU San Pablo

Festividad de San José

Marzo 2024

**Escuela Politécnica Superior
Universidad CEU San Pablo**

La inteligencia artificial en el análisis de datos: técnicas y aplicaciones en el uso de la voz

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Dirijase a CEDRO (Centro Español de Derechos Reprográficos, www.cedro.org) si necesita escanear algún fragmento de esta obra.

© Javier Tejedor Noguerales, 2024

© de la edición, Fundación Universitaria San Pablo CEU, 2024

CEU *Ediciones*

Julián Romea 18, 28003 Madrid

Teléfono: 91 514 05 73

Correo electrónico: ceuediciones@ceu.es

www.ceuediciones.es

Maquetación: Andrea Nieto Alonso (CEU *Ediciones*)

Depósito legal: M-6227-2024

Introducción

En primer lugar, quisiera dar las gracias tanto al director como al subdirector de la Escuela Politécnica Superior, por la confianza depositada en mí para la impartición de esta lección magistral en un día tan señalado como el patrón de San José de esta Escuela Politécnica Superior.

La temática de la misma está relacionada con la ciencia e ingeniería de datos, cuyos estudios de grado se imparten en la modalidad bilingüe en esta universidad, la Universidad CEU San Pablo y que paso a resumir en primer lugar.

Grado en Ciencia e Ingeniería de Datos

El grado en Ciencia e Ingeniería de Datos surge como respuesta a la creciente necesidad de las empresas de aplicar la inteligencia artificial con el fin de realizar una analítica de datos cada vez más avanzada. Aunque tiene puntos en común con otras titulaciones como Ingeniería Informática o Ingeniería Matemática, con esta titulación, el estudiante aprenderá a extraer conocimiento en cualquier ámbito científico y empresarial. Para ello, se combinan tres áreas de conocimiento como son la matemática, la estadística y la computación en un programa de cuatro años con un fuerte enfoque práctico. El objetivo final es que el estudiante aprenda las

técnicas y la aplicación de las herramientas informáticas más avanzadas de análisis de datos para proponer soluciones a problemas complejos y facilitar la toma de decisiones empresariales basadas en la información. De la misma forma, este grado oferta dos itinerarios formativos para que el estudiante se pueda especializar en el que más le guste. Uno enfocado a los sistemas de información empresarial, donde el estudiante aprenderá desde el análisis de datos usando sistemas web hasta aplicaciones móviles, y otro enfocado a la bioinformática, entendido este como la aplicación de técnicas de análisis de datos a las ciencias de la vida.

El creciente interés de las empresas en el análisis de datos está fundamentado en la llamada cuarta revolución industrial, la cual, indudablemente modificará la forma en la que vivimos, trabajamos y nos relacionamos unos con otros. Esta revolución industrial está basada en sistemas ciberfísicos que combinan infraestructura física con software, sensores, nanotecnología, tecnología digital y comunicaciones. En cuanto a las tecnologías que se enmarcan en esta cuarta revolución industrial se encuentra la inteligencia artificial, definida como la combinación de algoritmos que tienen como objetivo crear máquinas/procesos que realicen tareas de forma automática como los seres humanos (algunas ya incluso mejores que estos) y que, por tanto, presenten, cuanto menos, las mismas capacidades que estos. Un pilar básico de la inteligencia artificial es el anteriormente comentado análisis de datos, enfocado a la toma de decisiones a partir de algoritmos que intentan resolver una determinada tarea. Así, a partir del análisis de datos, podemos llegar a predecir el ganador de las próximas elecciones generales, estudiar patrones de cambio en pandemias, o estimar las posibles fugas de clientes de una compañía de seguros, por citar algunos ejemplos.

Inteligencia artificial y análisis de datos: técnicas

Para entrar un poco más en todo lo que nos ocupa, en el origen de todo, nos encontramos ante un problema complicado que no sabemos resolver y que, a través de la inteligencia artificial, podemos encontrar su solución. Esto, hoy en día, es aún magia para muchos de nosotros. Mi tarea, tras concluir esta lección es que no sea tal (o por lo menos tanto) para todos nosotros.

Desde un punto de vista práctico, la inteligencia artificial intenta abordar problemas relacionados con, entre otras muchas, las siguientes disciplinas:

- Reconocimiento de patrones. Así, a través del reconocimiento de patrones y, por ejemplo, una imagen, podemos llegar a reconocer la persona que está en ella, lo cual puede ser útil si queremos detectar el paso de personas no autorizadas a determinadas zonas de alta seguridad.
- Procesamiento del lenguaje natural. A través de la voz podemos reconocer las palabras que estoy pronunciando en este momento, para hacer aplicaciones que van desde el propio reconocimiento de voz, a sistemas de diálogo humano-máquina más avanzados.
- Predicción de eventos. A través de variables climatológicas, podemos llegar a predecir el tiempo que va a hacer un determinado día en una localización geográfica.
- Diagnóstico de enfermedades. A través de imágenes, voz, o electrocardiogramas podemos llegar a detectar patologías de una persona como por ejemplo tumores en cualquier parte del cuerpo, o enfermedades como el Alzheimer, el Parkinson y la apnea del sueño, entre otras muchas.

El hecho de usar máquinas (típicamente ordenadores de altas prestaciones) dota a los sistemas de inteligencia artificial de los siguientes puntos fuertes:

- El coste de replicación del sistema es mínimo. Una vez el sistema ha sido construido, la replicación es nula para la parte software (tan solo implicaría replicar el sistema en tantos ordenadores como queramos) y “prácticamente” nula para la parte hardware (tan solo haría falta la adquisición y puesta a punto de tantos ordenadores donde queramos replicar la parte software).
- Reproducibilidad. Un sistema computacional siempre proporciona las mismas salidas dado un mismo conjunto de entradas. Sin embargo, los seres humanos no somos tan consistentes, ya que puede haber situaciones personales que nos distraigan, podemos tener despistes a la hora de tomar una decisión,

etc. Por ponerles un ejemplo, y permítanme referirme al fútbol, como deporte rey de este país, el ser humano se puede llegar a influenciar por el desempeño de su equipo favorito en la jornada liguera anterior para tomar decisiones. Estas decisiones pueden ser tan graves como que un juez aumente la condena a una persona si su equipo favorito pierde, o dejarle absuelto si su equipo favorito golea al eterno rival. Evidentemente, los ordenadores son totalmente ajenos al desempeño de cualquier equipo en cualquier deporte que podamos pensar en este momento.

- Afectación por cansancio. Los ordenadores pueden trabajar 24 horas al día, 7 días a la semana sin cansarse. De la misma forma, si un ser humano, que está contratado en su empresa por 40 horas a la semana, típicamente 8 horas al día, intenta realizar un trabajo en su séptima hora laboral, este trabajo no será tan productivo como la primera hora laboral.

Tradicionalmente, existen dos tipos de técnicas o paradigmas para el análisis de datos. El primero de ellos está compuesto por estrategias basadas en conocimiento experto humano. En este, el humano adquiere un determinado conocimiento del problema a resolver y, una vez adquirido, realiza el sistema basado en conocimiento a partir de un mecanismo computacional de inferencia, tradicionalmente basado en reglas. Este paradigma, si bien puede funcionar aceptablemente bien en un entorno acotado y cerrado, tiene el principal problema de la adquisición del conocimiento, lo cual, para determinados problemas, puede llevar días o incluso meses. Por otra parte, el segundo paradigma es el denominado aprendizaje automático, también llamado *machine learning* o *data mining* y sus extensiones *deep learning* o *big data*. Si bien todas estas técnicas tienen ligeros matices (por ejemplo, el *deep learning* necesita de muchos datos para su construcción), todas tienen un mismo fin: que la máquina aprenda automáticamente a partir de los datos.

Centrándonos en este segundo paradigma, el de aprendizaje automático, podemos distinguir tres fases a la hora de diseñar, construir y evaluar el sistema:

- Fase de entrenamiento. En esta fase, el sistema, formado por uno o varios algoritmos de aprendizaje automático, construye el modelo a partir de los datos.
- Fase de desarrollo. En esta fase, estimamos la mejor configuración de parámetros para cuantos algoritmos formen parte del sistema.

- Fase de test. En esta fase, el sistema desarrollado se prueba y se analiza la bondad del mismo con las medidas de evaluación propias del área en la que se engloba el problema a resolver. Así, por ejemplo, si estamos desarrollando un sistema de reconocimiento de voz, nos interesa saber la tasa de error de palabra o *Word Error Rate* (WER). Esta medida es el porcentaje de error que tiene el sistema y se calcula a partir del número de inserciones, borrados y sustituciones en las palabras reconocidas. Una inserción se produce cuando el sistema detecta una palabra que no se dice en la realidad, un borrado cuando el sistema no detecta una palabra que sí se dice en la realidad, y una sustitución cuando el sistema confunde una palabra por otra. En el ejemplo donde la frase real es «Esta es la lección de inteligencia artificial» y el sistema de reconocimiento de voz reconoce «Esta es lección de una inteligencia artificioso», «la» sería un borrado, «una» sería una inserción y finalmente «artificioso» sería una sustitución (al reconocer «artificioso» cuando debería haber reconocido «artificial»). Por el contrario, si estamos desarrollando un sistema que detecta una posible enfermedad, nos interesaría evaluar la precisión o *Accuracy*, entendiendo esta como el porcentaje de sujetos que el sistema clasifica correctamente en función de si sufren o no dicha enfermedad.

Es necesario comentar en este punto que cada una de las tres fases mencionadas con anterioridad (entrenamiento, desarrollo y test), hace uso de un conjunto de datos diferente. Esto es crítico para el buen rendimiento del sistema cuando este esté en producción, es decir, se esté utilizando en la vida real. Así, por ejemplo, el uso de los mismos datos para entrenar y testear el sistema haría que el sistema memorizase (y por tanto no aprendiese) los datos, haciéndole prácticamente inservible cuando se enfrente a datos no vistos en producción. De la misma forma, el uso del mismo conjunto de datos para construir el modelo y estimar la configuración óptima de los parámetros podría conllevar un grave riesgo de sobreajuste (*overfitting*) del sistema a esos datos que mermaría notablemente su rendimiento en producción.

Inteligencia artificial en el procesado de voz

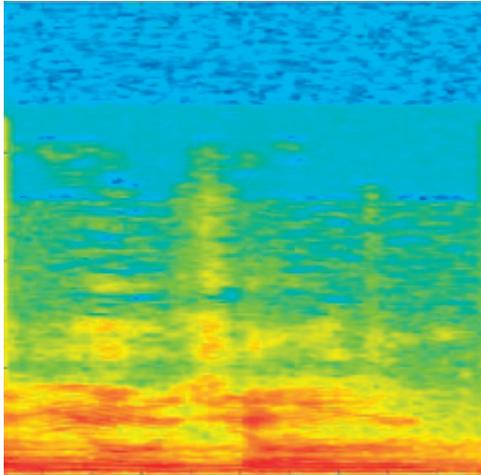
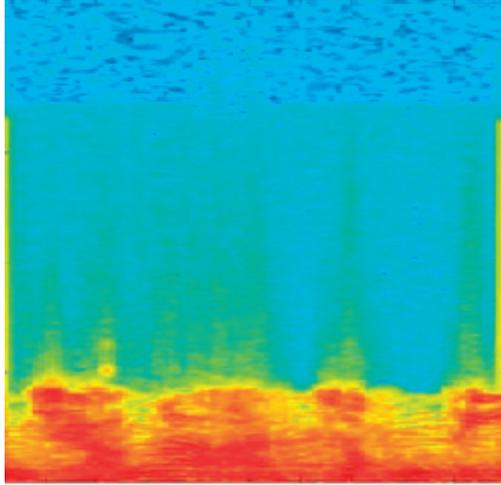
Una vez comentadas las técnicas generales aplicadas al análisis de datos, me centraré en la aplicación de la inteligencia artificial, y por ende del análisis de datos, a la forma de comunicación habitual del ser humano: la voz. La voz, entendida como el sonido que el aire expelido de los pulmones produce al salir de la laringe

haciendo que vibren las cuerdas vocales, es un elemento clave en nuestras vidas. Además, su marcado carácter no intrusivo, hacen de ella un elemento idóneo para su adquisición primero, y procesamiento después. Así, a través de la voz, se pueden realizar multitud de aplicaciones, entre las que destacamos:

- Reconocimiento de voz. Proceso mediante el cual se transcribe a texto el contenido acústico. Esta aplicación es muy útil para la realización de sistemas de diálogo humano-máquina donde se quiera dotar al mismo de una comunicación a través de la voz.
- Identificación de locutor. Proceso mediante el cual se sabe quién es la persona que está hablando. Esta aplicación es ampliamente usada por los cuerpos y fuerzas de seguridad en ámbitos forenses para identificar la voz de una determinada persona.
- Verificación de locutor. Proceso mediante el cual se verifica que la persona que está hablando es realmente quien dice ser. Este proceso es de vital importancia, por ejemplo, para sistemas de seguridad que hagan uso de la voz como rasgo biométrico. Y aquí es imprescindible remarcar la diferencia que hay entre la identificación y la verificación de locutor. Si bien ambas se enmarcan en la tecnología conocida como Reconocimiento de locutor, la identificación de locutor compara la voz del locutor en cuestión con las voces de todos los locutores que integran el sistema y, decide, en función de algoritmos de aprendizaje automático, qué modelo de locutor es el que corresponde al hablante en cuestión. Por otra parte, la verificación de locutor tan solo compara la voz del locutor objetivo con el modelo de la persona que dice ser dicho locutor, y a partir de los llamados umbrales de confianza, se verifica si realmente el locutor es quien dice ser.
- Reconocimiento de lenguaje. Proceso mediante el cual se reconoce el lenguaje en el que la persona está hablando. Este proceso puede ser útil cuando el diálogo que se establece entre humano y máquina es multilingüe, de tal forma que es necesario conocer el lenguaje en el que se está hablando para hacer uso de un reconocedor de voz del lenguaje a analizar.
- Síntesis de voz. Proceso mediante el cual se crea un archivo de voz con las palabras que dice un determinado texto. Este proceso, el más utilizado junto con el reconocimiento de voz por sistemas inteligentes que hacen uso de la voz, es

útil cuando no se puede disponer de voz pregrabada y se quiere construir un archivo de voz a reproducir al usuario. Imagínense el contexto de un sistema de diálogo humano-máquina en el que, a través de la voz, el usuario desea consultar la mejor ruta para ir de un sitio a otro usando diferentes modalidades de viaje: en coche, en transporte público o a pie. Tener un sistema con todas las posibles casuísticas de transporte para ir de cualquier sitio a cualquier sitio sería inviable desde el punto de vista de coste económico y de recursos. Por ello, la aplicación construye dinámicamente el archivo de voz a partir de las necesidades del usuario (el punto de origen y destino de su ruta junto con el medio de transporte elegido para ello) y a continuación se le reproduciría dicho archivo sintetizado. Es, en casos como estos, entre otros muchos, donde la síntesis de voz juega un papel fundamental. Afortunadamente, los sistemas de síntesis de voz en la actualidad, si bien no son tan naturales como la voz pregrabada, sí que permiten al usuario acceder a la información de una forma eficiente y cuasi-natural. Por ponerles un ejemplo, este archivo de voz contiene voz sintetizada, creada, por tanto, de forma artificial.

- Traducción de voz a lengua de signos. Para personas con graves deficiencias auditivas, pero que tienen conocimiento de lengua de signos en un determinado idioma, es posible realizar sistemas automáticos que traduzcan la voz a esa lengua de signos. De esta forma, se les puede hacer accesible la voz pronunciada por otra persona. Este tipo de aplicaciones integra varias tecnologías. En primer lugar, el reconocimiento de voz para transcribir el audio a texto. A continuación, traducción de texto a lengua de signos. Finalmente, un avatar se encarga de signar la información traducida a lengua de signos. En este vídeo pueden ver un sistema desarrollado en el marco de un proyecto de investigación en el que tuve la suerte de participar que traduce la voz del hablante a lengua de signos española.
- Detección de enfermedades a través de la voz. Hoy en día, es posible detectar enfermedades como el Parkinson, Alzheimer, apnea del sueño, por citar algunas, a través de la voz. Esto es posible gracias a que las características acústicas de personas que sufren estas enfermedades son diferentes a las que posee la voz de sujetos sanos. En estas dos imágenes se pueden observar diferentes patrones en las mismas que sugieren la aplicación de técnicas de inteligencia artificial. Esta imagen hace referencia a patrones identificados en el habla de una persona sana o sujeto control, mientras que esta segunda imagen hace referencia a patrones identificados en el habla de una persona con la enfermedad de Alzheimer.



Como se puede observar, hay cambios en los comportamientos de ambas imágenes (zonas más rojas en la primera imagen que en la segunda), que motivan el uso de técnicas de inteligencia artificial para poder llegar a predecir la enfermedad de Alzheimer a través de la voz.

Me centraré ahora en la primera de las aplicaciones mencionadas anteriormente, el reconocimiento de voz, con un poco de historia sobre la misma. Los orígenes de esta tecnología datan de 1950, cuando los laboratorios AT&T Bell (más conocidos como Bell Labs) crearon la primera máquina de reconocimiento de voz dependiente del locutor, *Audrey* (que solo funcionaba por tanto para la persona que entrenaba el sistema) y cuyo funcionamiento rondaba el 99% de precisión. Más tarde, en 1952 y a partir de una computadora analógica, se creó el primer sistema de reconocimiento de voz dependiente del locutor que reconocía el conjunto de dígitos del 0 al 9. Aunque este reconocedor no es tan general como el del idioma entero, es muy útil para el desarrollo de aplicaciones que contengan menús de navegación donde el usuario elige la opción que considere de una lista predefinida. En la década de los 60, se tomó conciencia de la complejidad de desarrollar una aplicación de reconocimiento de voz, y se empezaron a plantear aplicaciones que hacían uso de un vocabulario pequeño (no más de 200 palabras), y que eran dependientes del locutor. En 1970, la agencia de proyectos de investigación ARPA de la sección americana de defensa lanzó el primer sistema de reconocimiento de voz comercial. En concreto, la universidad americana Carnegie Mellon desarrolló *Harpy*, un sistema de reconocimiento de voz que necesitaba 50 ordenadores y que tenía un vocabulario de 1000 palabras (imagínense 50 ordenadores para reconocer, como máximo, 1000 palabras). Esto, a pesar de la envergadura del material hardware, supuso un hito hasta la fecha, ya que era capaz de no solo reconocer palabras sino también frases completas. Además, en esta década de los 70 es cuando se empiezan a hacer las primeras investigaciones en sistemas de reconocimiento de voz independientes del locutor, es decir, sistemas que pueden ser usados por cualquier hablante. Es en la década de los 80, concretamente en 1985, cuando se cambió por completo el paradigma en el que se basaban los sistemas de reconocimiento de voz. Si bien hasta entonces el reconocimiento de voz se basaba mayoritariamente en plantillas y reconocimiento de patrones basados en alineamiento temporal, es en 1985 cuando se crean los modelos probabilísticos llamados modelos ocultos de Markov para realizar el proceso de reconocimiento de voz. Todo esto fue posible gracias al crecimiento de los ordenadores personales, el apoyo de ARPA y los costos reducidos de aplicaciones comerciales. En esa misma década de los años 80, IBM lanzó su máquina de escribir que hacía uso de la voz llamada *Tangora*, y que era capaz de reconocer hasta 20.000 palabras diferentes. Es en los años 90 cuando empiezan a sonar con más fuerza aplicaciones comerciales que hacen uso del reconocimiento de voz. Así, por ejemplo, AT&T introdujo su *Voice*

Recognition Call Processing System en 1992 y Dragon Systems lanzó el primer producto de reconocimiento de voz para el usuario llamado *Dragon Dictate*, con su posterior actualización a *Dragon Naturally Speaking* en 1997, que podía reconocer la voz a un ritmo de 100 palabras por minuto. Tecnología que, por cierto, sigue funcionando hoy en día, tras su adquisición por parte de Microsoft en 2021. A principios del siglo XX, sistemas híbridos basados en redes neuronales y modelos ocultos de Markov eran los preferidos a la hora de diseñar sistemas de reconocimiento de voz. Hoy en día, y con el auge de la tecnología, donde cada vez hay ordenadores que procesan más rápido datos y datos de información, las redes neuronales profundas son la base de los sistemas de reconocimiento de voz catalogados como sistemas *end-to-end* ya que, a través de una única «función», se permite el paso de voz a texto.

Actualmente, los sistemas de reconocimiento de voz son los precursores de aplicaciones más avanzadas como la búsqueda en voz (*search on speech*) o la recuperación de documentos hablados (*spoken document retrieval*). Bajo estas aplicaciones, nos estamos refiriendo a la tarea de acceder a documentos de audio a través de una serie de consultas, como las que diariamente hacemos a buscadores de texto como Google, con el objetivo de obtener los documentos sonoros que son relevantes, es decir, que contienen las palabras clave de la consulta.

Otra de las grandes aplicaciones donde tiene cabida el reconocimiento de voz es, por ejemplo, en el ámbito de la domótica en las casas. De esta forma, incorporando la voz en las mismas, podemos por ejemplo pronunciar comandos que hagan las siguientes tareas del hogar: poner la lavadora, subir o bajar las persianas, encender o apagar la televisión, entre otras muchas.

De la misma forma, el uso de asistentes virtuales que hacen uso de la voz, como Alexa, Siri o el asistente virtual de Google, por nombrar algunos, hacen que nuestra vida pueda resultar simplemente más sencilla o divertida interactuando con ellos. Por ejemplo, además de poder controlar una casa domótica por voz, podemos hacer que dicho asistente virtual haga de meteorólogo, de cómico o simplemente nos diga qué hora es.

Conclusiones

Todo esto que acabo de presentar se enmarca en este grado en Ciencia e Ingeniería de datos, con el objetivo de formar profesionales que obtengan los conocimientos necesarios para desempeñar cualquiera de estos perfiles:

- Arquitecto del dato. Por hacer un símil con los estudios de Arquitectura, el arquitecto es el encargado de diseñar los planos del futuro edificio. Un arquitecto del dato, por tanto, se encarga de diseñar la infraestructura hardware y software donde se va a almacenar y procesar los datos.
- Ingeniero del dato. De nuevo sirviéndome de la comparación con Arquitectura, el obrero sería el que construye el edificio. Pues bien, el ingeniero del dato sería el «obrero del dato» y, por tanto, el que construiría la infraestructura hardware y software diseñada por el arquitecto y daría soporte a la misma.
- Científico de datos. Este perfil se encarga de desarrollar, típicamente, los algoritmos software que procesan los datos para dar la respuesta a un determinado problema y probarlos en la infraestructura creada con anterioridad.
- Analista de datos. Este perfil se encarga del análisis de los datos que engloba tareas como la limpieza, la recopilación, la visualización y la elaboración de informes a partir de los datos, trabajando por tanto en estrecha colaboración con el científico de datos.

Para concluir esta presentación, decidles que el análisis de datos y la inteligencia artificial están en un continuo crecimiento en la actualidad. A buen seguro estos campos nos van a proporcionar nuevas y fascinantes aplicaciones que bien con la señal de voz, o bien con otro tipo de señal fisiológica puedan, por qué no, llegar a detectar nuevas enfermedades que puedan surgir haciendo uso de la inteligencia artificial o hacer imposible saber si la persona que se encuentra al otro lado de la conversación es un humano o una máquina.

Reitero para finalizar mis agradecimientos a las personas que han depositado en mí la confianza para impartir esta lección magistral. Muchas gracias a todos.

JAVIER TEJEDOR NOGUERALES es profesor titular en la Universidad CEU San Pablo. Finalizó sus estudios en Ingeniería Informática en 2002 en la Universidad Autónoma de Madrid y recibió el título de Doctor en Ingeniería Informática y de Telecomunicación en 2009 por la Universidad Autónoma de Madrid. Desde 2003 hasta 2013, ha impartido docencia en la Universidad Autónoma de Madrid, en las titulaciones de Ingeniería Informática e Ingeniería de Telecomunicación. Desde 2014 hasta 2016 fue investigador post-doctoral en la Universidad de Alcalá. En 2015 y 2016 fue profesor asociado en ICAI, impartiendo docencia en el grado en Tecnología Industrial y desde 2016 es profesor en la Universidad CEU San Pablo, donde ha impartido docencia en los grados de Ingeniería de Sistemas de la Información e Ingeniería de Sistemas de Telecomunicación, así como en el máster en Ingeniería Biomédica. En la actualidad, imparte docencia en los grados de Ingeniería Biomédica y Ciencia e Ingeniería de Datos. Es el director del grado en Ciencia e Ingeniería de Datos en la Universidad CEU San Pablo y secretario del programa de doctorado en Ingeniería y Desarrollo Tecnológico en Aplicaciones Industriales, Biomédicas y Computacionales en la Universidad CEU San Pablo. También ha dirigido 21 trabajos fin de grado.

Su línea de investigación se centra en las tecnologías del habla, principalmente en la recuperación de información en contenidos acústicos. También realiza investigación en reconocimiento de patrones aplicado a señales de fibra óptica y a procesado de señal biomédica. Es el principal organizador de la evaluación internacional competitiva «Search on Speech», celebrada cada dos años desde 2012 dentro del congreso internacional «Iberspeech». Es autor/coautor de más de 90 publicaciones en revistas y congresos internacionales y nacionales, y ha participado en más de 30 proyectos financiados en convocatorias públicas y de financiación privada. En 2003, recibió el premio IMSERSO Infanta Cristina en el área de Nuevas tecnologías y ayudas técnicas y posee dos sexenios de investigación.