El peligro de la suplantación de identidad por medio de audio

Carlos Alberto Hernández Nava
Posgrado en Ciencias y Tecnologías de la Información, División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana – Iztapalapa
Eric Alfredo Rincón García
Pedro Lara Velázquez
Sergio Gerardo de los Cobos Silva
Miguel Angel Gutiérrez Andrade
Fabiola Margarita Martínez Licona
Alma Edith Martínez Licona
Departamento de Ingeniería Eléctrica, División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana – Iztapalapa.
Roman Anselmo Mora Gutiérrez
Departamento de Sistemas, División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana – Azcapotzalco.
Edwin Montes Orozco
Departamento de Matemáticas Aplicadas y Sistemas, División de Ciencias Naturales e Ingeniería, Universidad Autónoma Metropolitana – Cuajimalpa.

Foto proporcionada por los autores

Abstract
Biometric authentication has permeated daily life due to the continuous advancement of technology, which has allowed its inclusion in various services as well as in many everyday devices such as smartphones, laptops, or tablets. We must be aware of the danger posed by authentication through these means as identity theft attacks are a reality. This paper explains the vulnerabilities of automatic speaker verification authentication systems and why they are prone to attacks with audio generated for malicious purposes, as well as some necessary countermeasure approaches to achieve spoof audio detection and thus protect against identity theft.

Keywords
Audio spoof detection, Automatic speaker verification, Countermeasures, Neural networks.

Palabras clave
Detección de falsificación de audio, Verificación automática de locutor, Contramedidas, Redes neuronales.

Introducción

En los últimos años las redes sociales y la interacción que se da en ellas ha crecido bastante con respecto a años anteriores, por lo cual hoy en día se da en ellas un gran intercambio de información, tal como, imágenes de rostros, clips de voz, videos de movimiento humano de cuerpo completo, así como grandes cantidades de lenguaje natural, ya que los comentarios y reseñas son la base en la que se sostienen las aplicaciones de las redes sociales.

El avance continuo de las tecnologías y los dispositivos, así como la difusión de sensores de alto rendimiento en dispositivos al alcance de la población en general, ya sean celualres o laptops, trae consigo diversas aplicaciones benéficas para muchas y muy variadas áreas, pero este desarrollo ha tenido efectos secundarios como la pérdida paulatina de confianza en fotografías y videos, debido a los potentes programas que se usan para editarlos.

El progreso que ha tenido la tecnología de aprendizaje automático se debe en parte a la gran cantidad de información disponible en internet que se ha empleado para entrenar diferentes algoritmos de inteligencia artificial, los cuales son capaces de generar material falso tales como clips de voz, imágenes y videos de rostros humanos, así como textos en lenguaje natural que son muy parecidos a los genuinos.

La creación de este material audiovisual se está aplicando de manera beneficiosa a una amplia variedad de campos como el del entretenimiento, donde se ha utilizado en películas, como Rogue One parte de la saga de Star Wars, donde se utilizó para darle la apariencia de una actriz que había fallecido recientemente a otra actriz que interpretaría el mismo papel.

La parte negativa de esta capacidad de generar material digital es que puede ser utilizado para el fraude, la tergiversación y la falsificación, lo que plantea la amenaza de suplantación de identidad ya sea por audio, video o simplemente una imagen del rostro de la persona.

En un caso reportado en el 2019 por The Wall Street Journal (Stupp, 2019) se informó sobre un ataque en donde se utilizó software basado en inteligencia artificial, para hacerse pasar por el director ejecutivo de una empresa de energía con matriz

Figura 1. Métodos de autenticación biométrica más utilizados (adaptado de: https://pixabay.com/es/photos/hombre-rostro-reconocimiento-facial-5946820/, dedo-huella-dactilar-seguridad-2081169/, pódcast-audio-grabación-micrófono-2170045/).

en Alemania, mientras que otro director de la misma empresa con sede en Reino Unido pensó que estaba hablando por teléfono con su jefe, quien le solicitó realizar una transferencia de 220,000 euros a un supuesto proveedor húngaro, así que el pago fue realizado, lo anterior es claramente un ataque de suplantación de identidad utilizando únicamente audio.

La solución a esta problematica consiste en promover el uso de las tecnologías de autenticación, que ya tienen un lugar innegable en las aplicaciones y servicios actuales, por ejemplo, la huella dactilar, el escaneo de la palma de la mano, así como las huellas digitales, el escaneo de la retina, el reconocimiento del iris, el reconocimiento facial y la identificación por voz representan los métodos de autenticación biométrica más utilizados (Figura 1.).

En particular el reconocimiento de rostro y voz, ofrecen autenticación repetitiva y continua desde el canal de comunicación, o el flujo de la cámara, sin tener que realizar una solicitud adicional, estos flujos son una característica incluida en la mayoría de los dispositivos actuales ya sean celulares, tabletas o laptops.

Específicamente en el área del audio, encontramos que la verificación del hablante se usa a menudo en la telefonía, donde la voz es el único patrón biométrico posible, ya que no existe la posibilidad de contacto visual, por lo tanto, el sistema de verificación automática de locutor (ASV por sus siglas en inglés) es propenso a intentos maliciosos de autenticación, así podemos ver que el riesgo de suplantación en un ASV está definido y es claramente conocido.

Existen varias instituciones en México que hacen uso de este recurso para autenticación biométrica, el ejemplo más conocido es en los bancos, donde por medio de tu voz puedes autenticarte y realizar algunos procesos únicamente utilizando el teléfono.

Una vez dicho lo anterior es claro que existe un nicho para contramedidas a la amenaza antes mencionada, la detección de los medios audiovisuales falsos ha creado una atención generalizada, por lo que se están desarrollando aplicaciones que permiten detectar videos, fotografías y clips de voz no genuinos.

Bases de datos

El interés por lograr el reconocimiento

biométrico del habla y del hablante hizo evidente la necesidad de contar con una base de datos estandarizada para evaluar las contramedidas que iban surgiendo a lo largo de los años, por lo tanto para solventar esta necesidad, en 2015, el *National Institute of Informatics* inició dos desafíos vinculados a la clonación de medios y su identificación, con el fin de establecer una plataforma de evaluación unificada y criterios de evaluación para facilitar una comparación equitativa: el primero es el *Desafío de Conversión de Voz* y el segundo es el *Desafío ASVspoof*.

El Desafío de Conversión de Voz (Toda et al., 2016) es un evento que se celebra cada dos años desde 2016. En este desafío, los participantes reciben una base de datos y tienen la tarea de desarrollar convertidores de voz utilizando su propia tecnología, luego los organizadores evalúan el discurso convertido proporcionado por los participantes. La principal metodología de evaluación consiste en una prueba de comprensión auditiva, donde tanto los participantes como los organizadores valoran la semejanza del discurso creado de forma artificial con el original. Este desafío remarca la importancia de investigar los límites de las nuevas tecnologías para imitar la voz de una persona.

El segundo evento, conocido como Desafío ASVspoof (Wu et al., 2015), también se celebra cada dos años y surgió en 2013, realizando su primera edición en 2015. La versión más reciente de su base de datos correspondiente al año 2021 ya ha sido liberada.

De manera similar al Desafío de Conversión de Voz, se suministra a los participantes una base de datos estándar que contiene numerosos pares de audios genuinos y falsificados, los cuales pueden ser generados artificialmente y/o reproducciones de audios. El desafío consiste en que los participantes clasifiquen correctamente estos audios utilizando su propia tecnología, y los organizadores evalúan la precisión de detección de los modelos proporcionados por los participantes.

El gran éxito del *ASVspoof Challenge* 2015 de suplantación de identidad y contramedidas de verificación automática de hablantes, confirmó la necesidad de la detección de intentos de suplantación de identidad basados en síntesis de voz y técnicas de conversión de voz.

La segunda edición, es decir, el ASVspoof Challenge 2017, se centró en la tarea de detección automática de reproducción de audios, que se generan mediante grabaciones de la voz de un hablante y se reproducen en un sistema ASV en lugar de un discurso genuino, para lo cual se recopiló una gran cantidad de datos de reproducción de voz del mundo real en el idioma inglés.

Para la versión del *ASVspoof Challenge 2019*, se separó el corpus en dos escenarios:

- ASVspoof 2019 escenario LA: este escenario implica ataques de suplantación de identidad que se inyectan directamente en el sistema ASV, los ataques en este escenario son generados utilizando la última tecnología de síntesis de texto a voz (TTS) y tecnologías de conversión de voz (VC).

- ASVspoof 2019 escenario PA: para este escenario los datos de voz son capturados por un micrófono en un espacio físico reverberante, los ataques de repetición

de suplantación son grabaciones de habla auténtica que se capturan, subrepticiamente, y luego se vuelven a presentar al micrófono de un sistema ASV usando un dispositivo de reproducción.

Una de las bases, más conocidas pero que no corresponde a los desafíos es la Voice Spoofing Detection Corpus (VSDC), la cual se creó para ser un conjunto de datos estándar que contiene archivos de audio originales de diferentes entornos y micrófonos, así como archivos de audio falsificados generados a través de diversos entornos controlados, que simulan escenarios realistas.

Estos conjuntos de audios se pueden emplear como una herramienta para investigar los ataques de repetición convencionales, el impacto que tienen distintos tipos de micrófonos y entornos en la calidad del audio, o cómo la variabilidad en el rango vocal de una persona afecta el funcionamiento de un sistema controlado por voz. Estas bases de datos están cosntiudidas por un gran número de audios incluso la más pequeña entre ellas se compone de miles de audios, tal como se puede apreciar en la Tabla 1.

Base de datos	Número de audios
ASVspoof 2017 V2	18030
ASVspoof 2019 escenario LA	122299
ASVspoof 2019 escenario PA	218430
VSDC (Voice Spoofing Detection Corpus)	11772

Tabla 1. Bases de datos

Redes neuronales

Para protegerse de estas amenazas se han desarrollado diversos modelos, para la detección de audios que tienen como finalidad la suplantación de identidad, estos modelos pueden utilizar diferentes tipos de clasificadores para realizar la tarea deseada.

Uno de los enfoques más prometedores es utilizar redes neuronales, por lo tanto, se comenzó utilizando una sola red neuronal, para después con el tiempo ir aumentando las capacidades de estos modelos al utilizar redes cada vez más sofisticadas, así como combinaciones de varios tipos de redes neuronales.

Antes de describir algunos de los modelos que hacen uso de estas técnicas, es importante entender, ¿Qué es una red neuronal artiticial?, una definición muy acertada la da Haykin (Haykin, 1994, p. 24), la cual dice: “Combinación masivamente paralela de una unidad de procesamiento simple que puede adquirir conocimiento del entorno a través de un proceso de aprendizaje y almacenar el conocimiento en sus conexiones”, dichas conexiones pueden observarse en la Figura 2.

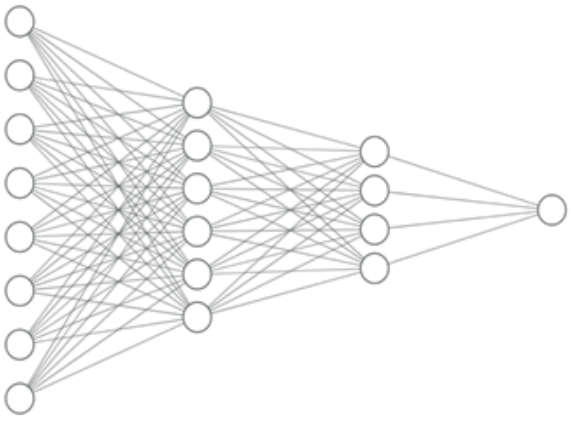

Figura 2. Red neuronal

Posteriormente se propondrían varios tipos de redes neuronales, una de las más destacadas son las redes neuronales con-

volucionales (CNN), que consisten en múltiples capas de filtros convolucinales tal como se observa en la Figura 3, y una de sus principales ventajas es que contienen un extractor de características compuesto de capas de convolución y capas de submuestreo, se ha demostrado que son muy eficaces y también las más utilizadas en diversas aplicaciones de visión y reconocimiento por computadora.

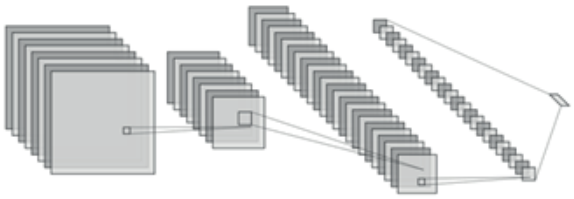

Figura 3. Red neuronal convolucional

También existen y son de uso común las redes neuronales recurrentes (RNN), que son una clase de aprendizaje profundo basadas en los trabajos de David Rumelhart en 1986. Las RNN son conocidas por su capacidad para procesar y obtener información de datos secuenciales. Por lo tanto, el análisis de vídeo, el procesamiento del lenguaje natural y el análisis de audio, se pueden realizar debido a las capacidades de las RNN, las cuales capturan las dependencias secuenciales y temporales.

Un tipo en particular de las RNN son las redes neuronales de memoria de corto y largo plazo (LSTM), cuyo propósito es solucionar el problema de las RNN asociado a la memoria de corto plazo: el desvanecimiento del gradiente, y su explosión. El principal objetivo de este tipo de redes es mejorar la memoria de la RNN de los eventos pasados entrenándola para que recuerde lo importante y olvide el resto de información que no le es relevante.

Contramedidas

A partir de la liberación de las bases de datos para los desafíos ASVspoof, se han hecho investigaciones de alto nivel con resultados muy favorables, como se mencionó uno de los enfoques es utilizar redes neuronales convolucionales (CNN) para resolver el problema de detección de ataques de reproducción (AR).

El éxito de CNN en tareas de clasificación y reconocimiento, como clasificación de video, clasificación de imágenes y reconocimiento facial fue una motivación poderosa para aplicar estos enfoques en tareas anti-spoofing de ASV, este enfoque puede extenderse a una variedad de tareas de clasificación de señales de audio, al representar la señal de entrada en un dominio de tiempo-frecuencia.

Surge la idea de combinar CNN y RNN y a pesar de no ser un enfoque nuevo los resultados siempre son favorables, muchos expertos encuentran que las CNN son buenas para extraer características en muchas tareas, en otras palabras, las características que son difíciles de diseñar a mano pueden ser abstraídas por CNN fácilmente, además, las RNN son buenas para capturar el mensaje secuencial del pasado hacia el presente para así encontrar sus dependencias.

Dado el poder computacional actual, que permite utilizar redes muy sofisticadas surge la duda sobre si es necesario implementar redes neuronales muy profundas o complicadas.

En (Pang & He, 2017), se mostró que no se requiere la implementación de redes

neuronales extremadamente profundas o complejas para detectar la suplantación de identidad.

Explicaron que con modelos sencillos es posible obtener resultados satisfactorios, su modelo consta de una capa de entrada, dos capas CNN 1-D, una capa de unidades recurrentes cerradas (GRU) y una capa final totalmente conectada tal como se aprecia en la Figura 4. Se obtiene un excelente rendimiento de este modelo con en el corpus compuesto por 28000 audios extraídos de la APSRD (Authentic and Playback Speaker Recognition Database).

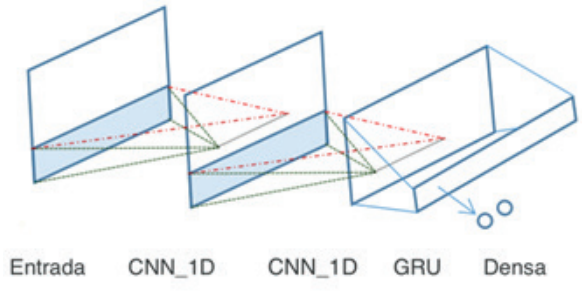

Figura 4. Arquitectura de un modelo sencillo para detección (adaptado de Pang and He 2017).

El otro enfoque es usar redes neuronales cada vez más profundas, pero esto puede causar el problema del desvanecimiento de gradiente. Para superar esta dificultad, surgieron las redes neuronales residuales (ResNet), que proporcionan un marco para entrenar redes más profundas.

Dado el éxito de ResNet en el reconocimiento de imágenes, (Chen et al., 2017) estudió su eficacia para la detección de falsificaciones. Los resultados con el conjunto de datos ASVspoof 2017 mostraron que ResNet tiene uno de los mejores rendimientos entre los sistemas de modelo único. De hecho, la fusión de modelos es una buena manera de mejorar el rendimiento del sistema, sin embargo, descubrieron que si se utilizan las mismas funciones para diferentes modelos fusionados, el sistema resultante difícilmente mejora.

Otro aspecto a tomar en cuenta es que hasta este punto no se ha mencionado si es realmente necesario obtener audio de alta calidad para lograr un ataque de audio, pero en (Lorenzo-Trueba et al., 2018) los investigadores examinaron la posibilidad de entrenar sistemas de suplantación de identidad utilizando exclusivamente datos de baja calidad.

Para ello, crearon un sistema de mejora del habla basado en redes generativas antagónicas (GAN), que eleva la calidad de los datos del habla disponibles públicamente en internet, logrando mejorar la claridad del habla sin comprometer la naturalidad de la voz en el audio.

La importancia de la calidad del audio se destaca debido al aumento en la popularidad de dispositivos controlados por voz (VCD), como Google Home, Amazon Alexa, Siri, entre otros, que han llevado a la automatización de electrodomésticos, dispositivos inteligentes, vehículos y servicios activados por voz. Dado que se ha demostrado que los ataques de voz no requieren audios de alta calidad, se puede concluir que estos dispositivos son susceptibles a ataques de reproducción de audio.

En un estudio de vulnerabilidades de VCDs llevado a cabo (Malik et al., 2020), se señala que estas repeticiones pueden ser aprovechadas en situaciones de múltiples saltos para acceder de manera maliciosa a dispositivos o nodos conectados a Internet de

las cosas (IoT). Por ejemplo, un dispositivo se utiliza para reproducir la voz del locutor, emitiendo una orden o comando a un segundo VCD, el cual la ejecuta sin verificar si la instrucción proviene genuinamente del locutor o es simplemente una repetición de la voz del mismo tal como se ve en la Figura5.

Figura 5. Escenarios de ataques de reproducción (adaptado de Malik et al. 2020).

Es evidente que las combinaciones de redes neuronales están presentes en muchos estudios recientes y sofisticados. En (Dua et al., 2021) los autores analizan el desempeño de diferentes arquitecturas que incluyen redes neuronales profundas (DNN), capas de redes de memoria a largo y corto plazo (LSTM), convolución temporal (TC), convolución espacial (SC), entre otras.

Cabe resaltar que en la UAM-Iztapalapa, como parte del trabajo de doctorado de Carlos Alberto Hernández Nava dentro del área de Optimización e inteligencia Artificial, del Departamenteo de Ingeniería Eléctrica, se ha trabajado en esta problemática y de dicho trabajo se han desprendido diversas propuestas, entre las cuales se ha destacado una al obtener excelentes resultados.

La propuesta consiste en un ensamblaje de varios modelos basados en redes neuronales, que trabajan en conjunto mediante una arquitectura modular y que los hace funcionar como contramedida.

Dentro del ensamble hay módulos formados por CNN que son capaces de diferenciar cuando un espectrograma proviene de un audio genuino o de uno falso a pesar de ser muy parecidos, tal como se aprecia en la Figura 6, lo que le ha permitido al sistema colocarse como una de las mejores técnicas de autentificación de audios actualmente.

El ensamble fue desarrollado apoyándose en la base de datos ASVspoof 2017 V2, la cual contiene esencialmente ataques de reproducción, es decir, son grabaciones de voz sobre las cuales se puede llevar a cabo algún proceso de modificación y repetirlas

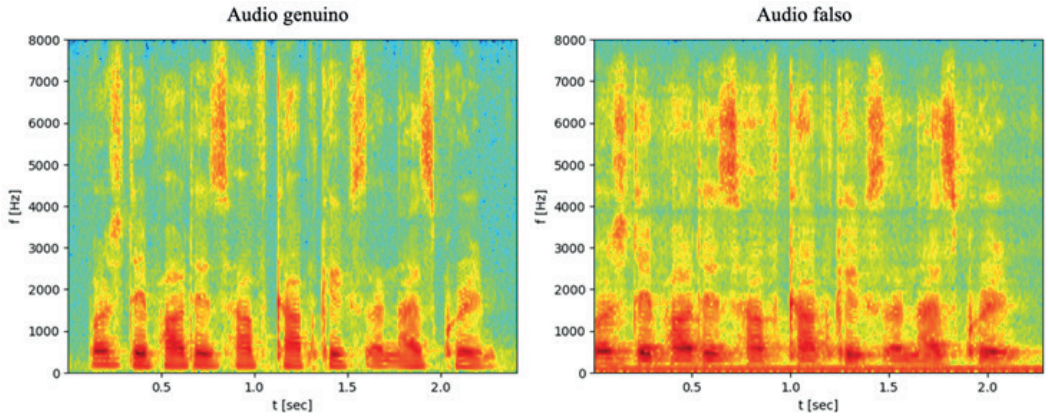

Figura 6. Espectrogramas de un audio genuino y un audio falso.

hacia un dispositivo o persona con la intención de engañarlo.

Esta base de datos como se mencionó más arriba consta de la cantidad de audios necesarios para realizar el diseño, entrenamiento y pruebas de los modelos que tienen como finalidad servir como contramedida.

El sistema es capaz de detectar si un audio es genuino o si es un audio falso, con un alto porcentaje de precisión del 96.46% de los 12000 audios, este es un excelente resultado como se puede apreciar más a detalle en (Hernández-Nava et al., 2023).

Una de las bondades de este desarrollo es que por la forma en que fue pensado es modular y escalable, por lo que puede servir como una base sólida para otros trabajos que deseen agregar módulos o funcionalidades al ensamble propuesto.

Conclusiones

Los sistemas que funcionan como contramedidas siempre deben estar abiertos a la mejora, ya que la capacidad de generación de audio es cada vez más sofisticada y por lo tanto los audios falsos o fraudulentos se hacen más difíciles de detectar.

El poder verificar y proteger la identidad de una persona mediante autenticación biométrica es una prioridad en un mundo moderno que se va digitalizando día a día y en el cual debemos ser precavidos con la información que llega a nosotros.

Finalmente es necesario recalcar la importancia de atender y trabajar en contramedidas que brinden la seguridad necesaria en un sistema de autenticación biométrica.

Referencias

Chen, Z., Xie, Z., Zhang, W., & Xu, X. (2017). ResNet and Model Fusion for Automatic Spoofing Detection. *18th Annual Conference Of The International Speech Communication Association (Interspeech 2017),* Vols 1-6: Situated Interaction, 102–106. https://doi.org/10.21437/Interspeech.2017-1085

Dua, M., Jain, C., & Kumar, S. (2021). LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems. *Journal of Ambient Intelligence and Humanized Computing, 13*, 1985–2000. https://doi.org/10.1007/s12652-021-02960-0

Haykin, S. (1994). *Neural Networks - A Comprehensive Foundation* (Second Edi). Pearson Education.

Hernández-Nava, C. A., Rincón-García, E. A., Lara-Velázquez, P., De-Los-Cobos-Silva, S. G., Gutiérrez-Andrade, M. A., & Mora-Gutiérrez, R. A. (2023). Voice spoofing detection using a neural networks assembly considering spectrograms and mel frequency cepstral coefficients. *PeerJ. Computer Science, 9*, e1740. https://doi.org/10.7717/peerj-cs.1740

Lorenzo-Trueba, J., Fang, F., Wang, X., Echizen, I., Yamagishi, J., & Kinnunen, T. (2018). Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. *Odyssey 2018 The Speaker and Language Recognition Workshop*. https://doi.org/10.21437/odyssey.2018-34

Malik, K. M., Javed, A., Malik, H., & Irtaza, A. (2020). A Light-Weight Replay

Detection Framework For Voice Controlled IoT Devices. *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, 14(5)*, 982–996. https://doi.org/10.1109/JSTSP.2020.2999828

Pang, W., & He, Q. (2017). A Simple Neural Network Based Countermeasure for Replay Attack. *Proceedings Of 2017 2nd International Conference On Communication And Information Systems*, 234–238. https://doi.org/10.1145/3158233.3159308

Stupp, C. (2019, August). Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *The Wall Street Journal*. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., & Sizov, A. (2015). *ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge*. https://doi.org/10.21437/Interspeech.2015-462