DOCTORAL THESIS

Deep learning applied to speech processing: Development of novel models and techniques

Submitted by

Roberto Andrés Carofilis Vasco

in fulfillment of the requirements for the Degree of

PhilosophiæDoctor (Ph.D.)

Doctoral Program: Production and Computer Engineering

A dissertation supervised by

Prof. Dr. Enrique Alegre Gutiérrez,

Prof. Dra. Laura Fernández Robles

León, September 2023

universidad
de león

TESIS DOCTORAL

Aprendizaje profundo aplicado al procesamiento de voz: Desarrollo de nuevos modelos y técnicas

desarrollada por

Roberto Andrés Carofilis Vasco

a fin de optar al grado de

Doctor por la Universidad de León

Programa de Doctorado: Ingeniería de Producción y Computación

Tesis doctoral dirigida por

Prof. Dr. Enrique Alegre Gutiérrez,

Prof. Dra. Laura Fernández Robles

León, septiembre de 2023

Abstract

This thesis proposes and evaluates new machine learning techniques and models for different tasks in the field of speech processing. It mainly addresses the identification of speakers, languages, and accents using several descriptor proposals based on different sound representations. In addition, it presents a new transfer learning technique based on a new descriptor, and two new architectures for deep learning models based on complementary audio representations.

The new transfer learning technique is based on a descriptor we call Grad-Transfer, which is based on the model interpretability method Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM generates a heat map of the most relevant zones in the input data according to their influence on a given model prediction. For the development of Grad-Transfer, we experimentally demonstrate, using Birch and k-means clustering algorithms, that the heat maps generated by the Grad-CAM method are able to store part of the knowledge acquired by a deep learning speech processing model fed by spectrograms during its training process. We exploited this capability of Grad-CAM to formulate a new technique that transfers knowledge from a pre-trained model to an untrained one, through the Grad-Transfer descriptor, which is responsible for summarizing and reusing such knowledge. Several Grad-Transfer-based models were evaluated for the accent identification task using the Voice Cloning Toolkit dataset. These models include Gaussian Naive Bayes, Support Vector Machines, and Passive Aggressive classifiers. Experimental results show an increase in performance of up to 23.58% in models fed by Grad-Transfer descriptors and spectrograms compared to models fed by spectrograms alone. This demonstrates the ability of Grad-Transfer to improve the performance of speech processing models and opens the door to new implementations for similar tasks.

On the other hand, new transfer learning approaches based on embedding generation models were evaluated. Embeddings are generated by machine learning models trained for a specific task on large datasets. By exploiting the knowledge already acquired, these models can be reused for new tasks where the amount of available data is small.

This thesis proposes a new architecture for deep learning models, called Mel and Wave Embeddings for Human Voice Tasks (MeWEHV), capable of generating robust embeddings for speech processing. MeWEHV combines embeddings generated by a pre-trained wave encoder model fed with raw audio and deep features extracted from Mel Frequency Cepstral Coefficients (MFCCs) using convolutional neural networks. We demonstrated the complementarity between the two representations and exploited it through neural layers specifically designed for their combination. We evaluated the performance of MeWEHV on three tasks: language identification, accent identification, and speaker identification. For the first task, we used the VoxForge and Common Language datasets. For the accent identification task, we used the Latin American Spanish Corpora and Common Voice datasets. Finally, for the speaker identification task, we used the VoxCeleb1 dataset and created YouSpeakers204, a new publicly available dataset for English speaker identification. YouSpeakers204 contains 19607 audio clips from 204 speakers with six different accents, allowing other researchers to work with a highly balanced dataset and build new models that are robust to multiple accents. This approach significantly improved the performance of the most advanced state-of-the-art models in all evaluated datasets, obtaining improvements of up to 88.27% in speaker identification, 14.86% in language identification, and 20.38% in accent identification. This was achieved at a low additional computational cost, with only 1.04M additional parameters, which represents between 0.33% and 1.09% more parameters than the pre-trained models used as a baseline.

In addition, a second architecture based on embedding generation models, called Squeeze-and-excitation for Embeddings Network (SaEENet), is proposed. SaEENet employs 1D depthwise separable convolution layers, GRU layers, and introduces, for the first time, the use of squeeze-and-excitation blocks for audio embedding weighting. The use of squeeze-and-excitation allows the model to assign a higher or lower relevance to each embedding generated from small audio segments, thus discarding information generated from voiceless segments or segments with non-relevant information. Furthermore, for the same architecture, we present experimental results using three different variations of squeeze-and-excitation blocks, identifying the most useful ones for the evaluated tasks. SaEENet outperforms MeWEHV and similar state-of-the-art models in the tasks of language identification, accent identification and speaker identification, achieving improvements of up to 0.90%, 1.41% and 4.01%, respectively, with 31.73% fewer trainable parameters than MEWHEV.

Overall, this thesis involves several advances in the areas of speaker, language, and accent identification, and proposes new techniques and models that use transfer learning to improve the performance of the state-of-the-art models evaluated.

Keywords: Languages identification, Accents identification, Speakers identification, Grad-CAM, Grad-Transfer, Speech processing, MeWEHV, SaEENet, Embeddings, YouSpeakers204

Resumen

Esta tesis propone y evalúa nuevas técnicas y modelos de aprendizaje automático en diferentes tareas dentro del campo del procesamiento del habla. Aborda principalmente la identificación de hablantes, idiomas y acentos, utilizando varias propuestas de descriptores basados en diversas representaciones del sonido. Además, presenta una nueva técnica de aprendizaje por transferencia basada en un nuevo descriptor, y dos nuevas arquitecturas para modelos de aprendizaje profundo basadas en representaciones de audio complementarias.

La nueva técnica de aprendizaje por transferencia se basa en un descriptor al que hemos denominado Grad-Transfer y que está basado en el método de interpretabilidad de modelos Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM genera un mapa de calor de las zonas más relevantes en los datos de entrada, según su influencia en una determinada predicción de un modelo. Para el desarrollo de Grad-Transfer demostramos experimentalmente, mediante los algoritmos de clustering Birch y k-means, que los mapas de calor generados por el método Grad-CAM son capaces de almacenar parte del conocimiento adquirido por un modelo de aprendizaje profundo de procesamiento del habla alimentado por espectrogramas, durante su proceso de entrenamiento. Aprovechamos esta capacidad de Grad-CAM para desarrollar una nueva técnica que transfiere conocimiento de un modelo preentrenado a uno sin entrenar, a través del descriptor Grad-Transfer encargado de resumir y reutilizar dicho conocimiento. Se evaluaron diversos modelos basados en Grad-Transfer para la tarea de identificación de acentos, usando el conjunto de datos Voice Cloning Toolkit. Entre estos modelos se encuentran los Gaussian Naive Bayes, Support Vector Machines, y clasificadores Passive Aggressive. Los resultados experimentales muestran un incremento de hasta el 23,58 % en el rendimiento en los modelos alimentados por descriptores Grad-Transfer y espectrogramas, en comparación de los modelos alimentados únicamente por espectrogramas. Esto demuestra que Grad-Transfer es capaz de mejorar el rendimiento de los modelos de procesamiento de voz y abre la puerta a nuevas implementaciones en tareas similares.

Por otra parte, se evaluaron nuevas aproximaciones de aprendizaje por transferencia basadas en modelos de generación de embeddings. Los embeddings son creados mediante modelos de aprendizaje automático entrenados en una tarea específica con grandes conjuntos de datos. Aprovechando los conocimientos ya adquiridos, estos modelos pueden reutilizarse en nuevas tareas en las que la cantidad de datos disponibles es reducida.

Esta tesis propone una nueva arquitectura para modelos de aprendizaje profundo, denominada Mel and Wave Embeddings for Human Voice Tasks (MeWEHV), capaz de generar embeddings robustos para el procesamiento del habla. MeWEHV combina los embeddings generados por un modelo wave encoder, preentrenado, alimentado por audio en bruto y características profundas extraídas de los Mel Frequency Cepstral Coefficients (MFCCs) mediante redes neuronales convolucionales. Su objetivo es demostrar experimentalmente la complementariedad entre ambas representaciones, y aprovecharla mediante capas neuronales específicamente diseñadas para su combinación. Evaluamos el rendimiento de MeWEHV en tres tareas: identificación de idiomas, identificación de acentos, e identificación de hablantes. Para la primera, utilizamos los conjuntos de datos VoxForge y Common Language. Para evaluar la tarea de identificación de acentos utilizamos los conjuntos de datos Latin American Spanish Corpora y Common Voice. Por último, para la tarea de identificación de hablantes utilizamos el conjunto de datos VoxCeleb1 y presentamos YouSpeakers204, un nuevo conjunto de datos puesto a disponibilidad del público para la identificación de hablantes de inglés. YouSpeakers204 contiene 19607 clips de audio de 204 personas que hablan con seis acentos diferentes, lo que permite a otros investigadores trabajar con un conjunto de datos altamente balanceado y crear nuevos modelos que sean robustos a múltiples acentos.

Nuestro enfoque permite aumentar significativamente el rendimiento de los modelos más avanzados del estado del arte, en todos los conjuntos de datos evaluados, consiguiendo una mejora de hasta el 88,27 % en identificación de hablantes, 14,86 % en identificación de idiomas, y 20,38 % en identificación de acentos. Necesitando para ello un bajo coste computacional adicional, al tener únicamente 1,04M parámetros adicionales, lo que representa entre un 0,33 % y 1,09 % más parámetros que los modelos preentrenados usados como baseline.

Adicionalmente, se propone una segunda arquitectura basada en modelos de generación de embeddings, llamada Squeeze-and-excitation for Embeddings Network (SaEENet). SaEENet emplea capas 1D depthwise separable convolutions, capas GRU, e introduce, por primera vez, el uso de bloques squeeze-and-excitation para la ponderación de embedddings de audio. El uso de squeeze-and-excitation permite al modelo asignar una relevancia mayor o menor a cada embedding generado a partir de pequeños segmentos de audio y descartar así la información generada a partir de segmentos sin voz o segmentos con información no relevante. Además, para esta misma arquitectura, presentamos resultados experimentales utilizando tres variaciones distintas de bloques squeeze-and-excitation, identificando, de esta forma, las más útiles para las tareas evaluadas. SaEENet supera a MeWEHV y a modelos similares del estado del arte en las tareas de identificación de idiomas, identificación de acentos e identificación de hablantes, logrando una mejora de hasta el 0,90 %, 1,41 % y 4,01 %, respectivamente, con un 31,73 % menos de parámetros entrenables que MEWHEV.

En conjunto, esta tesis presenta varios avances en las áreas de identificación de hablantes, idiomas y acentos, y propone nuevas técnicas y modelos que utilizan el aprendizaje por transferencia para mejorar el rendimiento de los modelos del estado del arte evaluados.

Palabras clave: Identificación de idiomas, Identificación de acentos, Identificación de hablantes, Grad-CAM, Grad-Transfer, Procesamiento del habla, MeWEHV, SaEENet, Embeddings, YouSpeakers204

Contents

List of Figures

List of Tables

Acknowledgements

I extend my sincere thanks to all the people who have supported me throughout the completion of this thesis.

I am deeply grateful to my supervisors, Dr. Enrique Alegre Gutiérrez and Dr. Laura Fernández Robles, for their unconditional support during my doctoral studies. Their patience, motivation, and knowledge have been invaluable.

I would like to thank Dr. Petr Motlicek for inviting me and supervising my research visit to the Idiap Research Institute in Martigny, Switzerland.

My sincere thanks to the Junta de Castilla y León for providing me with the grant to finance the recruitment of predoctoral research staff. I would also like to thank INCIBE and the European GRACE project for allowing me to collaborate with them and create tools that help the community.

From the bottom of my heart, I would like to thank my mom, and my siblings, who have been with me at all times, always being my support and my strength.

Finally, I would like to express my gratitude to my colleagues in the Group for Vision and Intelligent Systems at the University of Leon for their camaraderie and warmth.

Roberto Andrés Carofilis Vasco
León
October 15, 2023

Chapter 1

Introduction

1.1. Motivation

Speech-processing models have become increasingly important in various fields, such as law enforcement and cybersecurity. These models are used to fight against crimes such as child exploitation and human trafficking, as well as to identify suspects and provide evidence in criminal investigations. They can also be used for other applications, such as speech recognition in personal assistants, voice control systems, and language learning tools.

However, speech-processing models face several challenges being one of the main ones the lack of data. Since speech data are often limited and difficult to obtain, it can be challenging to train speech processing models that are accurate and robust on given tasks. In addition, speech processing models are often complex and require significant computing resources for training and running, which can be costly and time-consuming.

In this thesis, we address three tasks in speech processing: language identification, accent identification, and speaker identification.

1.1.1. Language identification

Language identification systems are essential tools in many different domains, from academia to industry and from homeland security to social network monitoring. These systems have a wide range of applications such as enhancing machine translation systems, filtering and classifying content in social networks, and facilitating communication in multilingual environments (Wang et al., 2022a; Nie et al., 2022).

One of the main challenges in developing effective language identification systems is data availability. Often, insufficient labeled data are available to train a model effectively, or the available data are unbalanced. Furthermore, even when a sufficient amount of labeled data is available, model training can require significant computational resources, which can be a bottleneck for many applications.

To address these challenges, there is a growing interest in developing language identification models that can achieve competitive performance with minimal data and computational resources (Shor et al., 2020; Conneau et al., 2021; Hsu et al., 2021; Chen et al., 2022).

These models have important implications for cybersecurity, as they can be used to detect and classify malicious content on the Web and other digital platforms. They can also support Law Enforcement Agencies (LEAs) in identifying suspects or monitoring online activities that pose a threat to national security.

The critical role of language identification systems in cybersecurity is underscored by the increasing use of social networks and online platforms. To effectively process and classify suspicious content in such environments, the automation provided by machine learning models has become essential (Fabien et al., 2021).

Interconnected model pipelines are often used in audio processing so that an Automatic Speech Recognition (ASR) model can generate the input of another model responsible, for example, for detecting suspicious language in the generated text. In these pipelines, language identification models are often crucial because the models in these pipelines are usually developed for specific languages, and their correct identification is critical (Fabien et al., 2021).

1.1.2. Accent identification

The ability to identify accents in speech is a critical problem in speech processing. Accents are influenced by the phonological, grammatical, and semantic aspects of the speaker's native language and can vary significantly from region to region and country to country (McArthur et al., 2018). Accent identification is a critical area of research with important applications in a wide range of fields.

Cybersecurity is a key application of accent identification. With the increasing reliance on authentication and voice recognition systems in security protocols, the ability to accurately identify accents has become essential. Identifying a speaker's country or region of origin can help identify potential threats and make tracking criminal activities across national borders easier (Zeng et al., 2019a; Najafian and Russell, 2020).

LEAs can benefit from accurate accent identification. By identifying the speaker's accent, investigators can narrow the list of potential suspects to people who match the same region of origin.

Accent identification can also help in fighting against child exploitation and other forms of delinquency. Children forced into sexual exploitation are often moved across borders and regions to avoid detection. Accent identification can help locate victims by identifying their country or region of origin. In addition, it can help identify the perpetrators of these crimes by tracking their accents and movements across borders.

In addition to its usefulness as a method for characterizing speaker traits, it has potential in a variety of fields, such as improving current ASR systems (Zeng et al., 2019a; Najafian and Russell, 2020). In general, accent is one of the most important factors influencing the performance of ASR systems, along with gender (Gupta and Mermelstein, 1982; Huang et al., 2001). Previous research has shown that accent is one of the biggest problems in creating variation-resistant ASR systems (Huang et al., 2004).

1.1.3. Speaker identification

The problem of speaker identification has important implications for various applications such as forensics, surveillance, and authentication. It involves determining the identity of an individual from his or her voice. Speaker identification systems have become increasingly popular due to the widespread use of voice-controlled devices, as well as the need for reliable and effective speaker recognition technology in various fields.

However, one of the main challenges in developing speaker identification systems is the need for large amounts of labeled data for training as well as significant computational resources. This makes difficult the application of these systems to real situations where resources are limited (Hsu et al., 2021; Chen et al., 2022). Therefore, there is a critical need to develop robust speaker identification models that can be trained with little data and require minimal computational resources while maintaining high accuracy (Nagrani et al., 2017; Garofolo, 1993; Cui et al., 2013; Pratap et al., 2020; Panayotov et al., 2015; Kahn et al., 2020).

To address these challenges, this thesis proposes two model architectures that can achieve competitive performance, while requiring minimal training data and computational resources. The proposed architectures leverage the latest advances in deep learning, including transfer learning, to improve the efficiency and accuracy of speaker identification systems.

In addition, this thesis also presents a new dataset that takes into account speaker accents, which is a crucial factor in the development of robust speaker identification models that can perform well in real-world scenarios.

A reliable and effective speaker identification system can help identify and track suspects, verify the identity of individuals and prevent crimes, as well as be able to identify missing persons (Fabien et al., 2021). In addition, the ability to develop these models with minimal data and computational resources makes them more accessible and cost-effective for organizations with limited resources.

Finally, a common challenge in language identification, accent identification, and speaker identification tasks is the lack of experimental configurations available in many of the publicly available datasets, which makes it difficult to fairly compare the results of new proposals with respect to the state-of-the-art.

For this reason, in this thesis, we propose a set of experimental setups for datasets that had no such configurations. We present the results achieved with our models using these experimental setups so that our results can be used as a baseline for future research.

1.2. Objectives

The main objective of this thesis is to develop new approaches and solutions for audio classification tasks, which outperform the state-of-the-art results at the time of publication. With this general objective in mind, we define the following particular objectives:

1. To create a large-scale dataset of labeled audio files to train machine learning models focused on accent identification, and identification of speakers with accents, to solve the problem of lack of balanced datasets and help the research community to develop robust models.

2. To develop a new transfer learning technique for speech classification that increases the performance of models trained from scratch.

3. To develop new architectures for deep learning models that take advantage of the benefits of using available transfer learning techniques while improving the performance of state-of-the-art models.

4. To study the complementarity between different types of acoustic representations to generate new types of robust audio representations.

5. To apply our research to a real-world problem focused on speaker profiling, aiming to extract relevant information from offenders and victims, as part of the European GRACE project.

1.3. Main Contributions

The main contributions of this dissertation are summarized as follows:

1. A novel feature extractor, called Grad-Transfer, which represents distinctive audio features, that combines information from the Convolutional Neural Networks (CNNs) based class discriminative localization technique Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) and from spectrograms.

2. A new method for transfer learning that uses Grad-Transfer, so that the method transfers knowledge from CNNs to classic machine learning algorithms.

3. A new speaker identification dataset, called YouSpeakers204, highly balanced in terms of speaker accents and gender, which was extracted from publicly available YouTube videos, and can be used for speaker, accent, and gender identification. This dataset was made publicly available, along with a proposed experimental setup, so that other researchers could compare their results in a fair way.

4. A new pipeline for the generation of rich audio embeddings, by merging multiple representations, which establishes a possible basis for new architectures that aim at improving large pre-trained models.

5. Two new model architectures, called Mel and Wave Embeddings for Human Voice Tasks (MeWEHV), and Squeeze-and-excitation for Embeddings Network (SaEENet), that achieve better results than the state-of-the-art models in speaker,

language, and accent identification tasks, requiring a relatively low number of trainable parameters.

6. For the first time in literature, we introduce squeeze-and-excitation neural layer blocks for weighing and filtering the information compressed in the embeddings generated by pre-trained models in speech processing tasks.

7. New publicly available experimental setups of the Voice Cloning Toolkit, Latin American Spanish Corpora, VoxForge, Common Language, Common Voice datasets, which can be used by other researchers to make a fair comparison of their results.

1.4. Publications and Research Results

1.4.1. Publications related to this Thesis

- Carofilis, Andrés, Fernández-Robles, Laura, Alegre, Enrique, & Fidalgo, Eduardo (2023). “Improvement of accent classification models through Grad-Transfer from Spectrograms and Gradient-weighted Class Activation Mapping" in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2859-2871, 2023, doi: 10.1109/TASLP.2023.3297961.

  Journal Impact Factor (JCR 2022): 5.4, Rank by Journal Impact Factor: Acoustics 3/31 (Q1), Engineering, Electrical & Electronic 61/275 (Q1).

- Carofilis, Andrés, Fernández-Robles, Laura, Alegre, Enrique, & Fidalgo, Eduardo (2023). “MeWEHV: Mel and Wave Embeddings for Human Voice Tasks" in IEEE Access, vol. 11, pp. 80089-80104, 2023, doi: 10.1109/ACCESS.2023.3300973.

  Journal Impact Factor (JCR 2022): 3.9, Rank by Journal Impact Factor: Computer Science, Information Systems 72/158 (Q2), Engineering, Electrical & Electronic 100/275 (Q2), Telecommunications 41/88 (Q2).

- Carofilis, Andrés, Fernández-Robles, Laura, Alegre, Enrique, & Fidalgo, Eduardo (2023). “Squeeze-and-excitation for embeddings weighting in speech classification tasks" in IEEE/ACM Transactions on Audio, Speech, and Language Processing. (Under Review).

  Journal Impact Factor (JCR 2022): 5.4, Rank by Journal Impact Factor: Acoustics 3/31 (Q1), Engineering, Electrical & Electronic 61/275 (Q1).

1.4.2. Attended Conferences

- Applications of Intelligent Systems, 2023, 23-27 January, Gran Canaria, Spain
- 2022 Language Recognition Evaluation (LRE) Workshop, 2023, 31 January, Online
- VIII Jornadas Nacionales de Investigación en Ciberseguridad (JNIC 2023), 2022, 21-23 June, Vigo, Spain
- ECCV 2022: 17th European Conference on Computer Vision, 2022, 23–27 October, Online
- VII Jornadas Nacionales de Investigación en Ciberseguridad (JNIC 2022), 2022, 27-29 June, Bilbao, Spain

1.4.3. Oral presentations at conferences

- Age and gender estimation from speech to support the detection of Child Sexual Abuse Material. Applications of Intelligent Systems, 2023, 23-27 January, Gran Canaria, Spain
- Presentation of the description of the system submitted by the University of León. 2022 Language Recognition Evaluation (LRE) Workshop, 2023, 31 January, Online.

1.4.4. Intellectual Property Registrations

- Patent: System, method, and program product for automatic accent classification in audio signals[1] (Registration ID: 202231062, Date: 13/12/2022, Status: Requested)

1.4.5. Awards and Grants

- This thesis was supported by the “Comunidad de Castilla y León”, through the scholarship order EDU/875/2021: "Ayudas para financiar la contratación predoctoral de personal investigador" (15/08/2021 - date of publication of this thesis)

1.4.6. Other Activities

Teaching Activities

- Direction of the Bachelor degree thesis: Santos A. (2020). Evaluación y mejora de la identificación del hablante utilizando deep learning (Assessment and improvement of speaker identification using deep learning). University of León.

[1] Spanish Patent and Trademark Office, published in Spanish.

- Direction of the Master's degree thesis: Castro M. (2023). Uso de Grad-CAM y redes convolucionales preentrenadas para detectar y clasificar poros presentes en superficies de piezas fabricadas con moldes de arena y cerámica (Use of Grad-CAM and pretrained convolutional networks to detect and classify pores present in surfaces of parts made with sand and ceramic molds). University of Vigo and University of León.

Projects

- "Acuerdo de Colaboración para la continuidad de los trabajos de un equipo de investigación aplicada en visión artificial y aprendizaje automático". Addendum 01 to the Framework Agreement between INCIBE (Spanish National Cybersecurity Institute) and the University of León.
- European GRACE Project: Global Response Against Child Exploitation. Grant agreement ID: 883341. DOI: 10.3030/883341.

Co-author in the following Registered Intellectual Property:

- Intellectual property: Application for the classification of fraudulent e-commerces[1] (Registration ID: 765-1019484, Date: 04/11/2022, Status: Accepted)
- Intellectual property: Application for the classification of Phishing URLs[1]. (Registration ID: LE-3-2021, Date: 05/01/2021, Status: Accepted)

1.4.7. Other Publications

- Martínez-Mendoza, Alicia, Sánchez-Paniagua, Manuel, Carofilis, Andrés, Jáñez-Martino, Francisco, Fidalgo, Eduardo & Alegre Enrique (2023). Applying Machine Learning to login URLs for phishing detection. VIII Jornadas Nacionales de Investigación en Ciberseguridad 2023, 487-488.
- Carofilis, Andrés, Chaves, Deisy, Martínez-Mendoza, Alicia, Fidalgo, Eduardo, González-Castro, Victor & Alegre, Enrique (2023). Impact of facial occlusions in age estimation algorithms for forensic applications. VIII Jornadas Nacionales de Investigación en Ciberseguridad 2023, 497-498.
- Biswas, Rubel, Del Río, Aitor, Vasco-Carofilis, Andrés, Swaroop, Guru, De Mata, Verónica & Alegre, Enrique. Image hashing based on frequency dominant neighborhood structure (2022). VII Jornadas Nacionales de Investigación en Ciberseguridad 2022, 294-295.
- Castaño, Felipe, Velasco, Javier, Vasco-Carofilis, Andrés, Fidalgo, Eduardo, Fernández, Luis & Azzopardi, George (2022). Evaluation of supervised learning models

using TCP traffic for the detection of botnets. VII Jornadas Nacionales de Investigación en Ciberseguridad 2022, 259-260.

- Blanco-Medina, Pablo, Fidalgo, Eduardo, Alegre, Enrique, Vasco-Carofilis, Roberto A.; Janez-Martino, Francisco & Fidalgo, Victor (2021). Detecting vulnerabilities in critical infrastructures by classifying exposed industrial control systems using deep learning. Applied Sciences, 11(1), 367.

1.4.8. Summer Schools

- REGINNA 4.0 Summer School: Deep Tech training with impact on entrepreneurship and innovaton. Organizer: University of Nova Gorica, Nova Gorica, Slovenia, 3-14th July 2023.

1.5. Thesis Structure

This chapter describes the structure of the doctoral thesis. This first introductory chapter focuses on motivating the work presented in this dissertation, its main objectives, and its original contributions. The remainder of this manuscript is organized as follows.

Chapter 2 contains a detailed review of state-of-the-art approaches related to the problems addressed in this thesis: language identification, accent identification, speaker identification, and related work on the proposed contributions. We also mention the main limitations of the methods reviewed and improvements that can be applied.

In Chapter 3, entitled "Improvement of accent classification models through Grad-Transfer from Spectrograms and Gradient-weighted Class Activation Mapping" the performance of popular CNNs architectures and Classical Machine Learning Algorithms (CMLAs) was evaluated for native English accent classification. Furthermore, we present Grad-Transfer, a novel descriptor based on the concatenation of a flattened spectrogram and dimensionality-reduced heat maps of Grad-CAMs. This descriptor takes advantage of the knowledge extracted by a CNN to be used as additional information to improve the distinctiveness of an audio descriptor and thus achieves higher performance with CMLA classifiers. The presented descriptor is especially useful in situations where a CMLA yields better performance than CNN models, thus further boosting the performance of the CMLA.

Chapter 4, entitled "MeWEHV: Mel and Wave Embeddings for Human Voice Tasks" presents a novel embedding enrichment procedure that combines the outputs of two models. On the one hand, a pre-trained embedding generation model from raw audio clips. On the other hand, the outputs of a neural network (NN), fed by the Mel Frequency Cepstral Coefficients (MFCCs) of the raw audios, which have among their advantages the capability of error reduction and robustness to noise. The main feature of MFCCs is that they focus on extracting relevant audio components to identify speech features and filtering other features, such as background noise, pitch, loudness, and emotion. The proposed architecture complements the high level of detail that the model exploits with the wave encoder, being this a non-imposed representation, and the extraction of relevant information through the MFCCs, as an imposed representation. For the correct complementarity of both types of representations, we designed an architecture capable of interacting with them through a set of layers, including LSTM layers and attention mechanisms. In this way, we managed to overcome the results obtained by other state-of-the-art models, at the same time requiring only a small number of trainable parameters. We also introduced a new manually labeled audio dataset called YouSpeakers204. Next, the chapter provides details of the empirical evaluation performed by us to verify our proposal.

Chapter 5, entitled "Squeeze-and-excitation for embeddings weighting in speech classification tasks", presents a Squeeze-and-excitation for Embeddings Network (SaEENet). SaEENet is a novel architecture that improves the state-of-the-art results in speaker identification, language identification and accent identification tasks. SaEENet is built using novel neural layers and several optimizations inspired by recent advances in other machine learning fields, such as the use of depthwise separable convolutions, and squeeze-and-excitation blocks, initially proposed in the field of image processing, and GRU layers, initially used in text processing. In the SaEENet model, we introduce a novel squeeze-and-excitation block that processes the stacked embeddings considering time as a dimension containing the target channels. Instead of weighting the relevance of the 2D channels of a convolutional network, SaEENet weights each 1D embedding according to its relevance. This allows the next layer to have context as to which embedding is more relevant, reducing the impact of embeddings generated from segments that do not contain speech or contain unnecessary information, and increasing the relevance of the segments that have information of interest to the model.

Chapter 6 summarizes the conclusions of this thesis and provides an outlook for possible future work lines to extend the presented work.

Chapter 2

State of the art

Multiple approaches have been applied in speech processing, which can be grouped into approaches using phonotactic modeling and approaches using acoustic modeling (Etman and Beex, 2015). Phonotactic modeling focuses on phoneme recognition and subsequent analysis, while acoustic modeling focuses on the spectral characteristics of sound waves or the raw waveform.

2.1. Phonotactic Modeling

All languages spoken by humans are composed of sets of phonemes, which are the basic theoretical units of sound that represent the minimum articulation of a vowel or a consonant and which are postulated to study spoken human language (Yallop and Fletcher, 2007)

Phonotactic modeling is characterized by processing an audio signal through its phonetic transcription, which is obtained through *phoneme recognizers* (Etman and Beex, 2015). A phoneme recognizer transcribes the voice into a sequence of known phonemes.

The phoneme recognizer produces phonetic sequences that are used to feed systems called language models (LMs), which are in charge of estimating a probability distribution model for each accent. LMs are models that assign probabilities to a sequence of words, such as the probability of a word appearing in a sentence given the previous set of words. Also, an LM can determine the probability that an n-gram (sequence of n words) belongs to a previously analyzed set of n-grams, which could represent a given accent (Martin and Jurafsky, 2018).

Most phonotactic modeling methods are based on statistical LMs, which are LMs that use traditional statistical techniques and linguistic rules to learn the probability distribution of words (Martin and Jurafsky, 2018).

Among the approaches that have used phonotactic modeling is the one presented by Kumpf and King (1996) who used an accent-dependent parallel phoneme recognizer to classify native Australian English speakers and foreign-accented speakers whose mother languages are Lebanese Arabic and South Vietnamese, using data extracted from the ANDOSL dataset (Vonwiller et al., 1995). The accuracies achieved are 85.3% and 76.6% for the classification of two and three accents, respectively.

Angkititrakul and Hansen (2006) implemented an accent classification system based on stochastic and parametric trajectory models using the CU-Accent corpus, an accent-sensitive word corpus. The corpus contains five English speaker groups with native American English and English spoken with Mandarin Chinese, French, Thai, and Turkish accents. This system achieved an accent classification accuracy of 90%.

Similar approaches have been used for language classification, such as (Safitri et al., 2016), in which they trained a statistical phonotactic model to classify the Indonesian languages Minangkabau, Sundanese, and Javanese. Using phone recognition followed by language modeling and parallel phone recognition followed by language modeling they achieved an accuracy of 77.4% and 75.94% respectively, using a non-public dataset.

Nie et al. (2022) proposed BERT-LID, based on a conjunction network for phoneme recognition and BERT with a linear output layer. They evaluated their proposal on the datasets AP20-OLR (Li et al., 2020), TAL_ASR, and a combination of the datasets THCHS-30 (Wang and Zhang, 2015) and TIMIT (Garofolo, 1993), achieving up to 5% improvement in audios of more than three seconds and 18% in audios of less than one second, with respect to models based on n-grams Support Vector Machines (SVM) and x-vectors.

Phonotactic modeling exhibits certain limitations. First, by focusing on patterns and linguistic features at the phoneme and word level, phonotactic modeling loses detailed acoustic information, such as intonation and rhythm, which are expressive aspects present in the voice signal. Additionally, this approach relies on precise transcriptions, which can be challenging in situations where accurate transcriptions are unavailable or when working with under-resourced languages. Furthermore, phonotactic models may struggle to capture individual variability among speakers, as they primarily focus on general phonetic patterns, resulting in a lack of personalization and adaptability to different voices and speaking styles. These limitations can affect the accuracy and robustness of phonotactic modeling in noisy environments or under adverse conditions Matejka (2009); Etman and Beex (2015). Due to this, although phonotactic modeling has been used in numerous studies, advances in deep learning have made speech processing research focus mostly on acoustic modeling (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022).

2.2. Acoustic Modeling

In general, acoustic models can be fed by raw audio waveforms (Hsu et al., 2021; Chen et al., 2022) or other spectral characteristics and representations extracted from them. Among these representations are spectrograms (Mulimani and Koolagudi, 2018; Zeng et al., 2019b; Sarthak et al., 2019), and Mel Frequency Cepstral Coefficients (MFCCs) (Lee and Jang, 2018; Ahmad et al., 2015), which can be competitive depending on the task and the dataset used, and in general both can obtain similar results (Meghanani et al., 2021).

Multiple investigations have used spectrograms as a representation of audio (Garain et al., 2021). A spectrogram is a result of calculating in frames the spectrum of a signal divided into windows; the result is a matrix containing information about the time, frequency, and energy of each instant (represented by color).

One of the ways to work with spectral characteristics of audio is through embeddings, which are representations capable of expressing statements with a variable number of observations as a single vector that retains most of the statement variations. There are several types of embeddings, such as i-vectors and x-vectors. Both have been widely used in audio classification tasks (Wang et al., 2020; Weninger et al., 2019; Krishna et al., 2019; Adeeba and Hussain, 2019).

The use of classification systems based on the attributes modeled by i-vectors has shown to be able to achieve superior results to those obtained by classic models based on phonotactic modeling, in tasks such as language recognition (Singer et al., 2012). A small set of speech attributes is sufficient for a complete characterization of spoken languages. Robust universal speech attribute detectors can be designed by sharing data between different languages, as shown by Siniscalchi et al. (2011).

Behravan et al. (2015) used i-vectors to define a common set of “universal” fundamental units that describe the manner and place of articulation as attributes of speech in all evaluated spoken accents. They evaluated the method on two datasets, the first dataset contains one native English group and another 7 groups of English speakers with a foreign accent, and the second dataset comprehends 7 groups of non-native speakers. They obtained an average detection cost of 5.02, and 6.30, respectively, representing an improvement of up to 8% and 15% over the result achieved with an approach based on the work of Siniscalchi et al. (2011).

Due to the improvement of Deep Neural Networks (DNNs) in recent years, a new type of embeddings has become popular, which are generated by extracting the outputs of a hidden layer of a pre-trained DNN. These embeddings are generated to store relevant information of an audio wave, to be later used in the learning of a new specific task. Therefore, it takes advantage of the modeling capabilities of a network trained with a large amount of data and reuses it with a small dataset.

A specific type of these embeddings are the x-vectors, which usually require the application of another model, such as Probabilistic Linear Discriminant Analysis (PLDA) after the generation of the embeddings, in order to determine which x-vectors are most similar to each other, and thus to determine the class of new x-vectors.

X-vectors have been used in related tasks. Wang et al. (2020) used x-vectors for speaker verification with Tagalog and Cantonese languages from the NIST SRE 2004-2010 datasets and achieved an equal error rate (EER) of 9.86%, and 2.96%, which represents an improvement of 15.4% and 34.7%, respectively, when comparing the results of x-vectors with the results of i-vectors.

Sun (2020) used x-vectors for language classification, with the NIST LRE07 dataset containing 14 languages and the Arabic dialect identification dataset containing 17 dialects, achieving an average accuracy of 75.64% and 71.85% with short audios respectively, representing an improvement of up to 10.5% and 9.8% when comparing x-vectors with i-vectors.

2.2.1. Approaches trained from scratch

In speech processing, several solutions and approaches that require training an acoustic model from scratch have been developed for individual tasks. Tang and Ghorbani (2003) made a comparison in accent identification between SVM and Hidden Markov Model (HMM), concluding that both approaches obtain similar performances and a similar speed of convergence.

The recent developments in the field of deep learning are focused on the use of models like DNNs, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs), as well as, in the fusion of these models with classic algorithms like SVM (Honnavalli and Shylaja, 2021; Zeng et al., 2019a; Ahmed et al., 2019; Zuo et al., 2015). Jiao et al. (2016) explored the combination of traditional and deep learning approaches for the task of automatic accent classification, using different combinations of models to classify 11 accents, and measuring the error rate by means of unweighted average recall (UAR). Among the results presented, SVM obtained a UAR of 45.1%, a model combining DNN+RNN achieved a UAR of 52.2% and a model combining DNN+RNN+SVM yielded a UAR of 55.8%.

With the evolution of CNNs and RNNs, much research in audio processing has focused on exploiting the representative capacity of CNNs and the ability to model temporal features of RNNs (Zuo et al., 2015). In language identification, architectures that combine CNN with RNN have been used, as in the case of the work of Bartz et al. (2017), where an architecture called Convolutional Recurrent Neural Network (CRNN) is proposed, specifically designed for learning from spectrograms. Later, Singh et al. (2020) applied CRNN and CNN to automatically classify accents, achieving 4.73% higher accuracy with CRNN than with CNN.

Another research where a CNN was used is the work of Zeng et al. (2019a), where a modified ResNet architecture was presented and applied to the multilabel classification of accents and speakers. With the purpose of accent classification, they achieved 89.67% accuracy on the VCTK dataset. This work combines the task of speaker classification and accent classification in the same model, so they do not separate the test and training set speakers. By having the same speakers in both subsets, the model is likely to learn to recognize the speakers, not the accent itself, and through speaker recognition, infer the accent.

Accent identification models have also been used as part of other systems, such as in the case of Ghangam et al. (2021), which focuses on creating an ASR system that is robust to speaker accent. To do this, they use language identification models and accent identification models to find the accent of an input audio, and subsequently analyze it with an ASR model specifically trained for that accent. The accent identification model used is composed of two Long Short-Term Memory (LSTM) layers of 200 neurons, and they tested it with Indian, Chinese and Malay accented audios, achieving a word error rate of 15.89, 26.05, 11.59, respectively. This is an improvement of up to 65.41%, 54.38%, and 72.87%, respectively, over other baseline models.

Wang et al. (2022b) presented a language identification system based on conformer layers, and a temporal pooling mechanism, which was tested on their own dataset with 65 languages and achieved an accuracy up to 4.27% higher than other approaches based on LSTM and transformers.

In speaker identification, Nassif et al. (2021) introduced the CASA-GMM-CNN model, in which they seek to clean a noisy audio through a Computational Auditory Scene Analysis (CASA). Then a classification of emotions is made by a Gaussian Mixture Models (GMM) and a CNN (GMM-CNN), and the output of both components feeds another GMM-CNN in charge of identifying the speaker. They tested their approach on SUSAS (Hansen and Bou-Ghazale, 1997), Arabic Emirati Speech Database (Shahin et al., 2020), RAVDESS (Livingstone and Russo, 2018), and Fluent Speech Commands (Lugosch et al., 2019) datasets, achieving up to 59.37% improvement in accuracy over other state-of-the-art works.

Nassif et al. (2022) presented another speaker identification model based on capsule networks, which is composed of two convolutional layers and one capsule layer, and it was compared using standard CNNs, random forests, GMM-DNNs, and SVMs as baseline models, on the Arabic Emirati Speech Database, SUSAS, and RAVDESS datasets. This model achieved improvements of up to 9.98%, 10.95%, and 9.81% accuracy, respectively, with respect to the best baseline model.

2.2.2. Transfer Learning Approaches

Transfer learning and domain transfer have been extensively studied in machine learning (Zhuang et al., 2021). Recent research related to transfer learning in speech processing has mainly focused on embeddings generation (Shor et al., 2020).

Wang et al. (2021b) presented an accent identification model generated from a pre-trained speech-to-text model, to which transfer learning was applied to be reused in their new task. To evaluate their proposal, they used the AP20-OLR dataset, achieving a reduction of up to 10.79% in the EER compared to other approaches based on x-vectors and i-vectors.

The most recent approaches are based on the use of pre-trained models and self-supervised learning methods for embeddings generation. These types of embeddings represent a position in an abstract multidimensional space, known as latent space, which encodes a meaningful internal representation of externally observed events. In these spaces, similar embeddings, or embeddings that have features in common, are close together, while less similar items are far apart (Kingma et al., 2014) One of these models is TRILL (Shor et al., 2020), which was trained using a subset of the AudioSet dataset (Gemmeke et al., 2017) and then evaluated across different domains using transfer learning and fine-tuning. The results achieved with TRILL were, in most cases, superior to those of the state of the art, and in other cases, close to them, being able to highlight its performance in speaker identification, with an accuracy of 17.9% on the VoxCeleb1 dataset (Nagrani et al., 2017), 94.1% for language identification on the VoxForge dataset (5.7% improvement) (MacLean, 2018), 91.2% for command identification on the Speech Commands dataset (Warden, 2018) (0.1% improvement), among others.

Other embedding generation models are the Wav2Vec2 (Baevski et al., 2020) model, which focused on English speech-to-text conversion, and XLSR-Wav2Vec2 (Conneau et al., 2021) model. XLSR-Wav2Vec2 is based on Wav2Vec2 but has been adapted for speech-to-text conversion in 53 languages, where the use of embeddings is useful to adapt the model to the different languages. To train the XLSR-Wav2Vec2 model, the MLS (Pratap et al., 2020), CommonVoice (Ardila et al., 2020), and the BABEL (Cui et al., 2013) datasets were used. The XLSR-Wav2Vec2 model was able to achieve a word error rate reduction of 72% compared to other published results on the Common Voice dataset, and 16% compared to the state-of-the-art results on BABEL.

Hsu et al. (2021) presented a new self-supervised approach for embedding generation based on BERT, called HuBERT. HuBERT uses an offline clustering step to provide aligned target labels for a BERT-like prediction loss. The HuBERT model matches or improves the performance of Wav2Vec2 on Librispeech (Panayotov et al., 2015) and Libri-Light (Kahn et al., 2020) datasets, achieving WER improvements of up to 19%.

A limitation of the HuBERT model is the fact that the Librispeech and Libri-Light datasets contain only English audio. Therefore, it is not demonstrated if the HuBERT model can work optimally in a multi-language environment.

Chen et al. (2022) presented the WavLM model extending the HuBERT framework for speech-to-text and denoising modeling, which enables pre-trained WavLM models to perform well on both speech-to-text and non-speech-to-text tasks. To achieve this, some WavLM inputs are noisy/overlapping speech simulations, and the expected outputs are the original speech labels. In addition, they optimized the model structure and training data of HuBERT and Wav2Vec2. The model was tested in the SUPERB Challenge (Yang et al., 2021) achieving an overall score 3.16% higher than HuBERT and 4.95% higher than Wav2Vec2.

Like HuBERT, WavLM was trained with the LibriSpeech and LibriLight datasets, but also used the GigaSpeech (Chen et al., 2021) and VoxPopuli (Wang et al., 2021a) datasets. WavLM has been trained taking into account multiple languages, thanks to the use of the VoxPopuli dataset, which contains 23 languages. Models based on Wav2Vec2, HuBERT and WavLM are fed with raw audio waveforms.

A common limitation in the embedding generation models mentioned above is the intensive use of resources to retrain them since they have a large number of parameters (318.42M in the case of Wav2Vec2-large, and 314.65 in the case of HuBERT-large). Consequently, a practical approach often employed is to reuse the embeddings generated by these models using transfer learning. This technique efficiently transfers the knowledge acquired during their initial training to new tasks. However, since the weights are not optimized for each specific new dataset, there is latent potential for improving model performance.

Deng et al. (2021) used the Wav2Vec2 model, in its large version, in the accent identification task and accented speech recognition. They worked with the Accented English Speech Recognition Challenge dataset (AESRC2020) (Shi et al., 2021), which contains 120000 training audio files, highly balanced among 8 accents, of which 6 are non-native. For accent identification, they propose a model that, by adding a set of fully connected layers, generates an accent prediction for each embedding generated by the model, and then combines the predictions into a single final prediction, instead of directly generating a single sentence-level prediction. They achieved up to 73.9% accuracy for accent identification, representing a 1.65% improvement over a model that generates a sentence-level prediction, and an 18.05% improvement over an i-vector-based approach.

Song et al. (2023) presented a model that uses WavLM to improve the speech enhancement task. Experiments were conducted on the DNS challenge dataset and on a simulation dataset. WavLM is used to generate embeddings of fixed-size windows, and that information is combined with information extracted from the original audio to generate new clean audio. The results show that the use of WavLM allows the improvement of speech enhancement systems, showing that the developed system achieves better performance than the other baseline models.

Apart from the models focused on speech processing, there are also models for general audio processing, such as the PANN model. The PANN model (Kong et al., 2020) was trained on the AudioSet dataset and evaluated using transfer learning and fine-tuning, in general content audio classification tasks. For environmental sound classification and audio taggings, PANN yielded accuracies of 94.7% and 96.0% on the ESC-50 (Piczak, 2015) and the MSoS (Kroos et al., 2019) datasets, respectively, surpassing the state-of-the-art results.

For acoustic scene classification, PANN was evaluated on the datasets DCASE-2019 (Mesaros et al., 2018) and DCASE-2018 (Fonseca et al., 2018), obtaining an accuracy of up to 76.4%, and 95.4%, respectively, in both cases lower than the state of the art. Whereas for music genre classification, PANN achieved an accuracy of 91.5% on the dataset GTZAN (Tzanetakis and Cook, 2002), lower than the state of the art. In all cases, the accuracy reported is higher than or close to the state-of-the-art results.

In addition to the well-known transfer learning methods, it is interesting to explore novel approaches that focus on specific cases where the available data is limited and unbalanced. Among the useful techniques in the development of new transfer learning methods is Gradient-weighted Class Activation Mapping (Grad-CAM).

Grad-CAM is a technique that weights data according to their relevance in the training process of a neural network. It is very popular in fields where it is necessary to have certainty about the reasons for the decisions made by machine learning models, such as, for example, the medical field, where it has been widely used because it allows the interpretability of the generated models and the consequent search for explanations of their outputs (Zhang et al., 2021; Moujahid et al., 2021; Kim et al., 2022; Nunnari et al., 2021). Grad-CAMs have also been used in the field of audio processing, specifically in the identification of acoustic events, allowing the interpretation of areas of interest in other visual representations of audio, such as MFCCs (Kim et al., 2021).

In this dissertation, we propose a new embedding enrichment method that makes use of the Grad-CAM algorithm in its process to extract features from the knowledge learned by a CNN and use it to improve the performance of another machine-learning model. To the best of our knowledge, there are no previous works that apply this approach.

2.2.3. Multi-representation Approaches

Approaches based on both reusing embedding generation models and models trained from scratch have demonstrated competitive performance in various audio processing tasks. However, all of them are based on a single representation of the original audio. Therefore, enrichment of the deep representations by another representation could improve the performance of such models.

Research on audio processing has significantly focused on the use of a single representation of the audio. Different representations and features extracted from an audio can be used at the same time to feed a model.

The combination of representations enables the extraction of complementary information from the original audio in a format that a machine learning model can easily process. This allows these models to outperform those that rely on a single representation. One example is FuzzyGCP (Garain et al., 2021), which is a model fed by eight types of representations generated from the original audios and which are joined into a single two-dimensional image. FuzzyGCP was evaluated for language identification on the datasets IIIT Hyderabad (Prahallad et al., 2012), IIT Madras (Baby et al., 2016), VoxForge, and MaSS (Boito et al., 2020), obtaining accuracies of 95%, 81.5%, 68%, and 98.7%, respectively. These results exceeded the ones obtained by other state-of-the-art approaches, such as PPRLM (Zissman and Singer, 1994), i-vector (Snyder et al., 2015) and x-vector (Snyder et al., 2018).

FuzzyGCP explores the combination of different audio representations and demonstrates superiority over classical approaches. However, it does not include raw audio representation as a possible input, thus not making use of the most recent developments in the field of speech processing.

FuzzyGCP does not make public the experimental setup with the training and test audios used, which makes it difficult to compare the obtained results.

Another model based on the combination of representations is the one proposed by Gao et al. (2022), in which they combined three types of audio representations that fed two models, one trained for acoustic scene classification and the other for general audio tagging. They use the DCASE 2018 Challenge dataset[1], achieving a mAP@3 of 93.26% in the audio tagging task and an accuracy of 72.48% in the acoustic scene classification task, outperforming the results of other state-of-the-art methods based on a single representation.

In this case, the combination of representations is done as ensemble models, where each individual model was trained autonomously with a different representation. The fusion of information is done in the output layer of the model through an information aggregation unit. Merging models through the model outputs has a limitation in that the information that can be shared in this way is limited compared to the information that could be obtained if more information-rich deep representations were interconnected.

Zhu et al. (2020) proposed a novel architecture fed by three types of representations, these representations fed two consecutive NNs. One network is responsible for identifying and filtering erroneously labeled training data so that they do not affect the training of the other network, thus avoiding data errors that may adversely affect the performance of the model. They tested their architecture in audio tagging with the FSDKaggle2018[2] and FSDKaggle2019[3] datasets, each one evaluated with a different metric, achieving a mAP@3 of 95.59%, and a label-weighted label-ranking average precision (lwlrap) of 0.7195 respectively, being, in both cases, competitive with the state-of-the-art methods.

This approach proved especially valuable when dealing with improperly filtered training data, which can impact the performance of models trained on relatively small datasets.

In this thesis, we propose a new architecture that enriches the embeddings generated by a pre-trained wave encoder model by combining it with embeddings extracted from MFCC representations through specialized neural layers in the architecture. It exploits the advantages of the benefits of embedding generation models and the combination of representations.

2.2.4. Squeeze-and-excitation Blocks

The embedding generation models generate an embedding for each segment of a fixed size in the original audio, this may imply that several of the generated embeddings contain the information of segments without voice or with excessive noise. A possible solution to reduce the impact of these embeddings is the use of Squeeze-and-excitation (SE) (Hu et al., 2020) blocks for embedding weighting.

SE blocks are model architecture units designed to improve the representational power of a network by allowing it to perform dynamic per-channel feature recalibration. Initially introduced as a mechanism to weight the channels generated by a convolutional layer, thus increasing or decreasing the values of each channel according to the relevance,

[1] http://dcase.community/challenge2018
[2] https://zenodo.org/record/2552860
[3] https://zenodo.org/record/3612637

for the next layer, of the data it contains (Hu et al., 2020).

SE blocks have been used mostly in image processing. Hu et al. (2020) showed an architecture called SENets, based on SE blocks, with which they achieved an improvement of up to 25% in the 2017 ILSVRC competition, concerning the winning model of 2016, using the ImageNet dataset (Russakovsky et al., 2015).

Recently, Zhang and Zhang (2022) presented SE-LPN-DPFF, a new model based on SE blocks, and evaluated it on the identification of SAR ships using the OpenSARShip (Huang et al., 2018), achieving an increase of up to 1.10% in accuracy over other state-of-the-art models.

Patacchiola et al. (2022) presented a new model based on an adaptation of the SE blocks, for few-shot image classification, improving by up to 1.5% in average accuracy using 18 datasets from the Visual Task Adaptation Benchmark (Dumoulin et al., 2021).

SE blocks have also been used in audio processing. Koluguri et al. (2022) presented TitaNet, an architecture for embedding generation in which SE blocks are applied to weight the channels of its convolutional layers. TitaNet was evaluated in the speaker verification task with the VoxCeleb1 dataset, achieving an EER of 0.68%, and in speaker diarization with the datasets AMI-MixHeadset, AMI-Lapel (Carletta et al., 2005) and CH109 (Canavan et al., 1997), achieving a diarization error rate (DER) of 1.73%, 1.99%, and 1.11%, respectively.

Huang et al. (2021) presented a hierarchical multi-embedding joint model, composed of two blocks, each with a Time Delay Neural Network (TDNN) and a RES2SETDNN model which is based on the same TDNN but adding a Res2Net type convolution and a squeeze-and-excitation block, and used a type of acoustic representation known as a Phone Posteriorgram (PPG). One block was fed with 40-dimensional PPGs and another block was fed with 120-dimensional PPGs. In addition, they used text-to-speech tools to generate synthetic speech data with each accent determined, thus achieving an accuracy of 83.63% on the AESRC2020 dataset.

In this thesis, we introduce a new architecture that adapts the SE blocks and uses them in the field of audio processing. Unlike TitaNet and other state-of-the-art models applying SE blocks in audio processing (Xu et al., 2020; Xue and Zhou, 2022; Rouvier and Bousquet, 2021; Yu et al., 2022), in SaEENet, SE blocks do not weight the channels generated by convolutional layers, but the information according to the relevance of the embeddings generated from different fixed-size segments of audio. To the best of our knowledge, this is the first time that squeeze-and-excitation blocks are used for embedding weighting in the context of speech processing.

Chapter 3

Improvement of accent classification models through Grad-Transfer from Spectrograms and Gradient-weighted Class Activation Mapping

This chapter focuses on the presentation of a new method called Grad-Transfer which is able to transfer knowledge from a deep learning model to a classical machine learning model used for accent identification.

Due to copyright issues, we have removed this chapter from the thesis. Here are the details of the published article:

Andrés Carofilis, Enrique Alegre, Eduardo Fidalgo, Laura Fernández-Robles, “Improvement of Accent Classification Models Through Grad-Transfer From Spectrograms and Gradient-Weighted Class Activation Mapping," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2859-2871, 2023, doi: 10.1109/TASLP.2023.3297961.

Chapter 4

MeWEHV: Mel and Wave Embeddings for Human Voice Tasks

This chapter presents a new deep learning architecture, called MeWEHV, for audio classification models, which uses complementary information from embeddings generated through pre-trained self-supervised learning models fed by raw audio and embeddings generated by CNNs fed by MFCCs. A new speaker identification dataset is also introduced which is highly gender-balanced and multi-accented.

Due to copyright issues, we have removed this chapter from the thesis. Here are the details of the published article:

Andrés Carofilis, Enrique Alegre, Eduardo Fidalgo, Laura Fernández-Robles, “MeWEHV: Mel and Wave Embeddings for Human Voice Tasks," in IEEE Access, vol. 11, pp. 80089-80104, 2023, doi: 10.1109/ACCESS.2023.3300973.

Chapter 5

Squeeze-and-excitation for embeddings weighting in speech classification tasks

This chapter presents a new architecture, called SaEENet, which is an improvement over MeWEHV, and introduces a novel mechanism for embedding weighting, which reduces the relevance of audio sections with missing or corrupted information. Other optimizations to the architecture are also introduced which allow us to reduce the total number of model parameters. Due to copyright issues, we have removed this chapter from the thesis. The related article is still in publication process.

Chapter 6

Conclusions and Outlook

6.1. Work Summary

This thesis evaluates and proposes new techniques and models in speech processing, using machine learning and deep learning. Three main tasks were addressed: (i) language identification, (ii) accent identification, and (iii) speaker identification. In recent years, there has been a great interest in the development of transfer learning techniques and their application in various models. This has helped to partially overcome known limitations, such as the lack of sufficiently large and balanced datasets to train complete models from scratch, and the need for large computational resources to execute the training phase of large models.

First, a new feature descriptor is proposed which we named Grad-Transfer. Grad-Transfer allows transferring a part of the knowledge acquired by a deep learning model during its training process, to a classical machine learning model, through a new implementation of the Grad-CAM model interpretability method.

In this work, we hypothesize that the position of elements in spectrograms plays a crucial role in determining the category of audio examples. We suggest that heatmaps generated from spectrograms using Grad-CAM, when used as inputs to a deep learning model for accent identification, can retain valuable information about the relevant features learned by the model, specifically in the position and shape of the identified points of interest in the heatmap. We demonstrated the stated hypothesis by using unsupervised learning methods and proposed a pipeline to generate Grad-Transfer descriptors, with which we managed to increase the performance of the evaluated classical machine learning models.

Another approach for applying transfer learning to speech processing is the creation of embedding generation models. Recent activity in the field is based on training large models with large datasets for a given task, with the purpose that the encoder layers learn to generate embeddings that summarize the most relevant information of an audio, and thus be able to reuse those embeddings in other tasks for which the model was not originally trained.

In this dissertation, we propose two new architectures for deep learning models, which we call Mel and Wave Embeddings for Human Voice Tasks (MeWEHV) and Squeeze-and-excitation for Embeddings Network (SaEENet). Both architectures are based on the initial hypothesis that a deep learning model can take advantage of the complementarity between the information contained in an imposed audio representation and a non-imposed representation and achieve better results than those obtained by a model fed only with one of the two types of representations.

MeWEHV takes as a baseline several of the most powerful state-of-the-art embedding generation models, which use raw audio waves as inputs and manage to increase their performance. The proposed architecture is composed of two branches of neural layers. The first one includes the encoder layers of a pre-trained model with its frozen weights, these layers generate a set of embeddings that are connected to subsequent layers in charge of processing and converting the set of embeddings into a single one. The second branch is fed by MFCCs that pass through a set of neural layers to generate a second embedding. The embeddings from the first and second branches are combined and used to generate the desired task outputs. MeWEHV was shown to outperform the state-of-the-art models used as a baseline, fed by raw audio, and a model fed only by MFCCs, in the tasks of language identification, accent identification, and speaker identification while requiring a relatively small number of new parameters concerning the baseline models.

SaEENet represents an improvement over MeWEHV and adds multiple new features, such as the use of depthwise separable convolution layers, GRU layers, and the introduction, for the first time in the context of audio processing, of an adaptation of squeeze-and-excitation blocks for embedding weighting. Squeeze-and-excitation blocks were initially proposed in the field of computer vision for weighting the channels of a feature map generated by convolutional layers. Thus, these blocks can increase or decrease the relevance that the elements of each channel should have to improve the overall performance of the model. In audio processing, embedding generation models divide any input audio into small segments and generate an embedding of each segment. These segments can be generated from noisy information or voiceless segments. For SaEENet, we use the new implementation of squeeze-and-excitation blocks to automatically increase or decrease the relevance of each of these embeddings. With the introduced changes, we managed to outperform MeWEHV on three different datasets with the tasks of language, accent, and speaker identification, requiring fewer trainable parameters.

In addition to this, a new audio dataset, called YouSpeakers204, was created using several audios extracted from public YouTube videos and made publicly available. The particularity of YouSpeakers204 is that it includes, among the metadata of each audio, a pseudonymized identifier of the speaker, its gender, and its accent. The accent was obtained by performing a manual search of each speaker's information and it corresponds to their country of origin. YouSpeakers204 is highly balanced in terms of gender and accent and allows the creation of robust speaker identification models for multiple accents, age identification models, and accent identification models. It contains 19607 audio clips, 204 speakers and 6 accents.

In the rest of the chapter, we present the main conclusions of this work and potential future lines of research.

6.2. General Conclusions

This dissertation provided successful solutions to multiple tasks within the field of speech processing, introducing new techniques and models. Some specific conclusions that can be drawn from this research are:

1. Using the unsupervised learning techniques K-means and Birch, it was demonstrated that the Grad-CAM model interpretability method can extract and store part of the knowledge acquired by a model during its training phase for the accent identification task. For this purpose, a dimensional reduction was applied to a set of heatmaps generated with Grad-CAM from a pre-trained accent classification model, and it was verified, using K-means and Birch, that in most cases the clusters generated correspond to the accent to which each example belongs. Hence, it could be inferred that the heatmaps contain part of the knowledge acquired by the pre-trained model during its training process.

2. The proposed new feature descriptor, called Grad-Transfer, is able to transfer the knowledge acquired by a pre-trained deep learning model to a classical machine learning model, leveraging the proven knowledge extraction and storage capabilities of Grad-CAM. The application of Grad-Transfer to the task of accent identification proved to be able to increase the performance of classical models. In this way, classical models fed by Grad-Transfer along with spectrograms achieved better results than classical models fed by spectrograms alone and also than baseline deep learning models fed by raw audio.

3. Using the new deep learning architecture Mel and Wave Embeddings for Human Voice Tasks (MeWEHV), which focuses on the generation of rich embeddings, it was experimentally demonstrated that there is a complementarity between the information contained in embeddings generated from multiple self-supervised learning models trained on raw audio and embeddings generated from MFCCs. MeWEHV is composed of two inputs, on the one hand, raw audio waves, which pass through a pre-trained model and a set of layers trained from scratch, responsible for generating an embedding E_1, and on the other hand, MFCCs generated from the same audio, which feed a set of layers trained from scratch, generating an embedding E_2. The embeddings E_1 and E_2 are merged and feed a set of classification layers. MeWEHV exhibited superior performance in language identification, accent identification, and speaker identification tasks compared to baseline models fed only by raw audio or fed only by MFCCs, requiring only a 0.21% increase in the number of total parameters with respect to the largest baseline model evaluated.

4. In the same line, a second architecture for deep learning models, called Squeeze-and-excitation for Embeddings Network (SaEENet), demonstrated multiple improvements over the MeWEHV architecture. SaEENet implements several new fea-

tures with respect to MeWEHV, including the reduction of the number of trainable parameters through the use of GRU layers, Depthwise Separable Convolutions, and, for the first time, the presentation of squeeze-and-excitation neuron blocks applied to the weighting of embeddings. By using the proposed squeeze-and-excitation blocks, SaEENet can increase or decrease the relevance of embeddings generated from different segments of the original audio, thus reducing the impact of embeddings generated from noisy or irrelevant information. SaEENet proved to be able to achieve better results on language, accent, and speaker identification tasks than MeWEHV and other state-of-the-art models on several of the datasets evaluated. SaEENet reduced the number of trainable parameters by 31.73% with respect to the reference model MeWEHV.

5. Through a comparative analysis of different types of squeeze-and-excitation blocks, it was determined which squeeze-and-excitation block variations achieve the best results, taking into account the branch of the SaEENet architecture in which they are used. The variations evaluated are: Spatial Squeeze and Channel Excitation (cSE), Channel Squeeze and Spatial Excitation Block (sSE), and Spatial and Channel Squeeze & Excitation (scSE). It was experimentally concluded that for the weighting of feature maps of a set of Depthwise Separable Convolutions layers, the best results were achieved with scSE blocks, and for the weighting of embeddings the best results were achieved with cSE blocks. These results can serve as a reference for future research related to the enrichment of audio embeddings.

6.3. Open Problems and Future Work

In this section, we summarize the main lines of work that remain open, as well as potential future work.

First, we discuss the presentation of the Grad-Transfer descriptor and its use for knowledge transfer between a deep learning model and a classical machine learning model. Hence, we can extract the following potential future work:

1. The implementation of Grad-Transfer focused on the task of accent identification, but a potential line of future research is the application of the proposal to other audio processing tasks, since the premise of Grad-Transfer is to take advantage of spectrogram regions derived from heat maps. In general, this principle can be applied to any spectrogram-fed task.

2. It would also be interesting to compare the performance of Grad-Transfer with other audio representations, such as MFCCs, and to evaluate its impact on the performance of different models.

Secondly, the MeWEHV and SaEENet architectures for voice processing models were presented. Next, the open research directions based on these architectures are mentioned:

1. Both MeWEHV and SaEENet demonstrate that combining raw audio with MFCCs in the same architecture yields better results than both representations used separately. In future research, an alternative to be explored is the use of other imposed representations in addition to MFCCs. Moreover, both architectures are composed of two branches of neural layers, each one fed by a different type of representation. Based on this, it is interesting to work with architectures composed of more than two branches, where different combinations between several types of audio representations, such as spectrograms and cochleograms, can be evaluated.

2. SaEENet introduces a new implementation of squeeze-and-excitation blocks for embedding weighting. A potential future research line is the creation of "cross squeeze-and-excitation" blocks that weigh the embeddings of a model, taking as inputs not only the previous layer in its corresponding branch but also information from the other branch of the model. A future cross squeeze-and-excitation block could infer which information is redundant in different representations of the inputs and reduce their relevance. Similarly, it could increase the relevance of information that is only in one corresponding branch for subsequent layers.

3. In order to build the SaEENet architecture, several variations were tested, including the use of bidirectional LSTMs, and multi-head self-attention layers, but in our experimentation they did not outperform the results obtained with the proposed architecture. A possible future line of research would be to perform a deeper analysis with other architectures based on these layers, or other new state-of-the-art layers.

4. Taking into account the results of MeWEHV and SaEENet, it is verified that there is a complementarity between the embeddings generated by a pre-trained model from raw audio and the MFCCs extracted from the same audios, since the results obtained using both representations outperform the results achieved using each representation separately. A future line of research would consist of determining the causes of this complementarity, establishing what is the missing information in both representations, in order to be able to propose, for example, a new representation that contains enough information to compete with the results presented in this thesis, using only one type of input.

Bibliography

Adeeba, F. and Hussain, S. (2019). Native language identification in very short utterances using bidirectional long short-term memory network. *IEEE Access*, 7:17098–17110.

Ahmad, K. S., Thosar, A. S., Nirmal, J. H., and Pande, V. S. (2015). A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network. In *Eighth International Conference on Advances in Pattern Recognition*, pages 1–6. IEEE.

Ahmed, A., Tangri, P., Panda, A., Ramani, D., and Karmakar, S. (2019). VFNet: A convolutional architecture for accent classification. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4. IEEE.

Angkititrakul, P. and Hansen, J. H. L. (2006). Advances in phone-based modeling for automatic accent classification. *IEEE Transactions on Speech and Audio Processing*, 14(2):634–646.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Baby, A., Thomas, A. L., Nishanthi, N., Consortium, T., et al. (2016). Resources for indian languages. In *Proceedings of Text, Speech and Dialogue*.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). Wav2Vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*.

Bartz, C., Herold, T., Yang, H., and Meinel, C. (2017). Language identification using deep convolutional recurrent neural networks. In *International Conference on Neural Information Processing*, pages 880–889. Springer.

Behravan, H., Hautamäki, V., Siniscalchi, S. M., Kinnunen, T., and Lee, C.-H. (2015). I-vector modeling of speech attributes for automatic foreign accent recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1):29–41.

Boito, M. Z., Havard, W., Garnerin, M., Ferrand, É. L., and Besacier, L. (2020). Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6486–6493. European Language Resources Association.

Canavan, A., Graff, D., and Zipperlen, G. (1997). Callhome american english speech.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2005). The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., Jin, M., Khudanpur, S., Watanabe, S., Zhao, S., Zou, W., Li, X., Yao, X., Wang, Y., You, Z., and Yan, Z. (2021). Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In Hermansky, H., Cernocký, H., Burget, L., Lamel, L., Scharenborg, O., and Motlícek, P., editors, *INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*, pages 3670–3674. ISCA.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., and Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In Hermansky, H., Cernocký, H., Burget, L., Lamel, L., Scharenborg, O., and Motlícek, P., editors, *INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*, pages 2426–2430. ISCA.

Cui, J., Cui, X., Ramabhadran, B., Kim, J., Kingsbury, B., Mamou, J., Mangu, L., Picheny, M., Sainath, T. N., and Sethy, A. (2013). Developing speech recognition systems for corpus indexing under the IARPA babel program. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013*, pages 6753–6757. IEEE.

Deng, K., Cao, S., and Ma, L. (2021). Improving accent identification and accented speech recognition under a framework of self-supervised learning. In Hermansky, H., Cernocký, H., Burget, L., Lamel, L., Scharenborg, O., and Motlícek, P., editors, *INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*, pages 1504–1508. ISCA.

Dumoulin, V., Houlsby, N., Evci, U., Zhai, X., Goroshin, R., Gelly, S., and Larochelle, H. (2021). Comparing transfer and meta learning approaches on a unified few-shot classification benchmark. *CoRR*, abs/2104.02638.

Etman, A. and Beex, A. L. (2015). Language and dialect identification: A survey. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, pages 220–231. IEEE.

Fabien, M., Parida, S., Motlícek, P., Zhu, D., Krishnan, A., and Nguyen, H. H. (2021). ROXANNE research platform: Automate criminal investigations. In Hermansky, H., Cernocký, H., Burget, L., Lamel, L., Scharenborg, O., and Motlícek, P., editors, *INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*, pages 962–964. ISCA.

Fonseca, E., Plakal, M., Font, F., Ellis, D. P., Favory, X., Pons, J., and Serra, X. (2018). General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline. In *Scenes and Events 2018 Workshop (DCASE2018)*, pages 69–73.

Gao, L., Xu, K., Wang, H., and Peng, Y. (2022). Multi-representation knowledge distillation for audio classification. *Multimedia Tools and Applications*, 81(4):5089–5112.

Garain, A., Singh, P. K., and Sarkar, R. (2021). FuzzyGCP: A deep learning architecture for automatic spoken language identification from speech signals. *Expert Systems with Applications*, 168:114416.

Garofolo, J. S. (1993). TIMIT acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017*, pages 776–780. IEEE.

Ghangam, S., Whitenack, D., and Nemecek, J. (2021). Dyn-Asr: Compact, multilingual speech recognition via spoken language and accent identification. In *7th IEEE World Forum on Internet of Things, WF-IoT 2021*, pages 830–835. IEEE.

Gupta, V. and Mermelstein, P. (1982). Effects of speaker accent on the performance of a speaker-independent, isolated-word recognizer. *The Journal of the Acoustical Society of America*, 71(6):1581–1587.

Hansen, J. H. L. and Bou-Ghazale, S. E. (1997). Getting started with SUSAS: a speech under simulated and actual stress database. In Kokkinakis, G., Fakotakis, N., and Dermatas, E., editors, *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997*. ISCA.

Honnavalli, D. and Shylaja, S. (2021). Supervised machine learning model for accent recognition in english speech using sequential mfcc features. In *Advances in Artificial Intelligence and Data Engineering*, pages 55–66. Springer.

Hsu, W., Bolte, B., Tsai, Y. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2020). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023.

Huang, C., Chen, T., and Chang, E. (2004). Accent issues in large vocabulary continuous speech recognition. *International Journal of Speech Technology*, 7(2-3):141–153.

Huang, C., Chen, T., Li, S., Chang, E., and Zhou, J. (2001). Analysis of speaker variability. In *Seventh European Conference on Speech Communication and Technology*.

Huang, H., Xiang, X., Yang, Y., Ma, R., and Qian, Y. (2021). AISpeech-SJTU Accent identification system for the accented english speech recognition challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021*, pages 6254–6258. IEEE.

Huang, L., Liu, B., Li, B., Guo, W., Yu, W., Zhang, Z., and Yu, W. (2018). OpenSARShip: A dataset dedicated to Sentinel-1 ship interpretation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(1):195–208.

Jiao, Y., Tu, M., Berisha, V., and Liss, J. M. (2016). Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features. In *INTERSPEECH*, pages 2388–2392.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., and Dupoux, E. (2020). Libri-Light: A benchmark for ASR with limited or no supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7669–7673. IEEE.

Kim, J., Oh, J., and Heo, T.-Y. (2021). Acoustic scene classification and visualization of beehive sounds using machine learning algorithms and Grad-CAM. *Mathematical Problems in Engineering*, 2021.

Kim, J.-K., Jung, S., Park, J., and Han, S. W. (2022). Arrhythmia detection model using modified DenseNet for comprehensible Grad-CAM visualization. *Biomedical Signal Processing and Control*, 73:103408.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3581–3589.

Koluguri, N. R., Park, T., and Ginsburg, B. (2022). TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022*, pages 8102–8106. IEEE.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). PANNs: large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Krishna, G. R., Krishnan, R., and Mittal, V. (2019). An automated system for regional nativity identification of indian speakers from english speech. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4. IEEE.

Kroos, C., Bones, O., Cao, Y., Harris, L., Jackson, P. J. B., Davies, W. J., Wang, W., Cox, T. J., and Plumbley, M. D. (2019). Generalisation in environmental sound classification: The 'making sense of sounds' data set and challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019*, pages 8082–8086. IEEE.

Kumpf, K. and King, R. W. (1996). Automatic accent classification of foreign accented australian english speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1740–1743. IEEE.

Lee, R. A. and Jang, J. R. (2018). A syllable structure approach to spoken language recognition. In Dutoit, T., Martín-Vide, C., and Pironkov, G., editors, *Statistical Language and Speech Processing - 6th International Conference, SLSP 2018*, volume 11171 of *Lecture Notes in Computer Science*, pages 56–66. Springer.

Li, Z., Zhao, M., Hong, Q., Li, L., Tang, Z., Wang, D., Song, L., and Yang, C. (2020). AP20-OLR challenge: Three tasks and their baselines. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2020*, pages 550–555. IEEE.

Livingstone, S. R. and Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, 13(5):e0196391.

Lugosch, L., Ravanelli, M., Ignoto, P., Tomar, V. S., and Bengio, Y. (2019). Speech model pre-training for end-to-end spoken language understanding. In Kubin, G. and Kacic, Z., editors, *INTERSPEECH 2019, 20th Annual Conference of the International Speech Communication Association*, pages 814–818. ISCA.

MacLean, K. (2018). Voxforge.

Martin, J. H. and Jurafsky, D. (2018). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Pearson/Prentice Hall Upper Saddle River, 3 edition.

Matejka, P. (2009). *Fonotaktické a Akustické RozpoznáVání Jazyků ; phonotactic and acoustic Language Recognition.* PhD thesis, Brno University of Technology, Czech Republic.

McArthur, T., Lam-McArthur, J., and Fontaine, L. (2018). *The Oxford companion to the English language.* Oxford University Press, 2 edition.

Meghanani, A., S., A. C., and Ramakrishnan, A. G. (2021). An exploration of log-mel spectrogram and MFCC features for alzheimer's dementia recognition from spontaneous speech. In *IEEE Spoken Language Technology Workshop, SLT 2021*, pages 670–677. IEEE.

Mesaros, A., Heittola, T., and Virtanen, T. (2018). A multi-device dataset for urban acoustic scene classification. In *Scenes and Events 2018 Workshop (DCASE2018)*, pages 9–13.

Moujahid, H., Cherradi, B., Al-Sarem, M., Bahatti, L., Eljialy, B. A., Alsaeedi, A., and Saeed, F. (2021). Combining CNN and Grad-CAM for COVID-19 disease prediction and visual explanation. *Intelligent Automation & Soft Computing*, 32(2):723–745.

Mulimani, M. and Koolagudi, S. G. (2018). Acoustic event classification using spectrogram features. In *TENCON 2018 - 2018 IEEE Region 10 Conference*, pages 1460–1464. IEEE.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. In Lacerda, F., editor, *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, pages 2616–2620. ISCA.

Najafian, M. and Russell, M. (2020). Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Speech Communication*.

Nassif, A. B., Shahin, I., Elnagar, A., Velayudhan, D., Alhudhaif, A., and Polat, K. (2022). Emotional speaker identification using a novel capsule nets model. *Expert Systems with Applications*, 193:116469.

Nassif, A. B., Shahin, I., Hamsa, S., Nemmour, N., and Hirose, K. (2021). CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions. *Applied Soft Computing*, 103:107141.

Nie, Y., Zhao, J., Zhang, W., and Bai, J. (2022). BERT-LID: leveraging BERT to improve spoken language identification. In Lee, K. A., Lee, H., Lu, Y., and Dong, M., editors, *13th International Symposium on Chinese Spoken Language Processing, ISCSLP 2022*, pages 384–388. IEEE.

Nunnari, F., Kadir, M. A., and Sonntag, D. (2021). On the overlap between Grad-CAM saliency maps and explainable visual features in skin cancer images. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 241–253. Springer.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210. IEEE.

Patacchiola, M., Bronskill, J. F., Shysheya, A., Hofmann, K., Nowozin, S., and Turner, R. E. (2022). Contextual squeeze-and-excitation for efficient few-shot image classification. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Piczak, K. J. (2015). ESC: dataset for environmental sound classification. In Zhou, X., Smeaton, A. F., Tian, Q., Bulterman, D. C. A., Shen, H. T., Mayer-Patel, K., and Yan, S., editors, *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM.

Prahallad, K., Elluru, N. K., Keri, V., S, R., and Black, A. W. (2012). The IIIT-H indic speech databases. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, pages 2546–2549. ISCA.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. (2020). MLS: A large-scale multilingual dataset for speech research. In Meng, H., Xu, B., and Zheng, T. F., editors, *INTERSPEECH 2020, 21st Annual Conference of the International Speech Communication Association*, pages 2757–2761. ISCA.

Rouvier, M. and Bousquet, P. (2021). Studying squeeze-and-excitation used in CNN for speaker verification. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021*, pages 1110–1115. IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252.

Safitri, N. E., Zahra, A., and Adriani, M. (2016). Spoken language identification with phonotactics methods on minangkabau, sundanese, and javanese languages. *Procedia Computer Science*, 81:182–187.

Sarthak, Shukla, S., and Mittal, G. (2019). Spoken language identification using ConvNets. In Chatzigiannakis, I., de Ruyter, B. E. R., and Mavrommati, I., editors, *Ambient Intelligence - 15th European Conference*, volume 11912 of *Lecture Notes in Computer Science*, pages 252–265. Springer.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Shahin, I., Nassif, A. B., and Hamsa, S. (2020). Novel cascaded gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments. *Neural Computing and Applications*, 32(7):2575–2587.

Shi, X., Yu, F., Lu, Y., Liang, Y., Feng, Q., Wang, D., Qian, Y., and Xie, L. (2021). The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021*, pages 6918–6922. IEEE.

Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., de Chaumont Quitry, F., Tagliasacchi, M., Shavitt, I., Emanuel, D., and Haviv, Y. (2020). Towards learning a universal non-semantic representation of speech. In Meng, H., Xu, B., and Zheng, T. F., editors, *INTERSPEECH 2020*, pages 140–144. ISCA.

Singer, E., Torres-Carrasquillo, P. A., Reynolds, D. A., McCree, A., Richardson, F., Dehak, N., and Sturim, D. E. (2012). The MITLL NIST LRE 2011 language recognition system. In *Odyssey 2012: The Speaker and Language Recognition Workshop*, pages 209–215. ISCA.

Singh, U., Gupta, A., Bisharad, D., and Arif, W. (2020). Foreign accent classification using deep neural nets. *Journal of Intelligent & Fuzzy Systems*, 38(5):6347–6352.

Siniscalchi, S. M., Lyu, D.-C., Svendsen, T., and Lee, C.-H. (2011). Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):875–887.

Snyder, D., Garcia-Romero, D., and Povey, D. (2015). Time delay deep neural network-based universal background models for speaker recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 92–97. IEEE.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5329–5333. IEEE.

Song, H., Chen, S., Chen, Z., Wu, Y., Yoshioka, T., Tang, M., Shin, J. W., and Liu, S. (2023). Exploring WavLM on speech enhancement. In *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*, pages 451–457. IEEE.

Sun, L. (2020). Spoken language identification with deep temporal neural network and multi-levels discriminative cues. In *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pages 153–157. IEEE.

Tang, H. and Ghorbani, A. A. (2003). Accent classification using support vector machine and hidden markov model. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 629–631. Springer.

Tzanetakis, G. and Cook, P. R. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

Vonwiller, J., Rogers, I., Cleirigh, C., and Lewis, W. (1995). Speaker and material selection for the australian national database of spoken language. *Journal of Quantitative Linguistics*, 2(3):177–211.

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J. M., and Dupoux, E. (2021a). Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 993–1003. Association for Computational Linguistics.

Wang, D., Ye, S., Hu, X., Li, S., and Xu, X. (2021b). An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model. In Hermansky, H., Cernocký, H., Burget, L., Lamel, L., Scharenborg, O., and Motlícek, P., editors, *INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*, pages 3266–3270. ISCA.

Wang, D. and Zhang, X. (2015). THCHS-30 : A free chinese speech corpus. *CoRR*, abs/1512.01882.

Wang, Q., Yu, Y., Pelecanos, J., Huang, Y., and Lopez-Moreno, I. (2022a). Attentive temporal pooling for conformer-based streaming language identification in long-form speech. In Zheng, T. F., editor, *Odyssey 2022: The Speaker and Language Recognition Workshop*, pages 255–262. ISCA.

Wang, Q., Yu, Y., Pelecanos, J., Huang, Y., and Lopez-Moreno, I. (2022b). Attentive temporal pooling for conformer-based streaming language identification in long-form speech. In *Odyssey 2022: The Speaker and Language Recognition Workshop*, pages 255–262. ISCA.

Wang, S., Yang, Y., Wu, Z., Qian, Y., and Yu, K. (2020). Data augmentation using deep generative models for embedding based speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2598–2609.

Warden, P. (2018). Speech Commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209.

Weninger, F., Sun, Y., Park, J., Willett, D., and Zhan, P. (2019). Deep learning based mandarin accent identification for accent robust asr. In *INTERSPEECH*, pages 510–514.

Xu, J., Wang, X., Feng, B., and Liu, W. (2020). Deep multi-metric learning for text-independent speaker verification. *Neurocomputing*, 410:394–400.

Xue, J. and Zhou, H. (2022). Physiological-physical feature fusion for automatic voice spoofing detection. *Frontiers of Computer Science*, 17(2):172318.

Yallop, C. and Fletcher, J. (2007). *An introduction to phonetics and phonology*. Blackwell Publishers.

Yang, S., Chi, P., Chuang, Y., Lai, C. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G., Huang, T., Tseng, W., Lee, K., Liu, D., Huang, Z., Dong, S., Li, S., Watanabe, S., Mohamed, A., and Lee, H. (2021). SUPERB: speech processing universal performance benchmark. In Hermansky, H., Cernocký, H., Burget, L., Lamel, L., Scharenborg, O., and Motlícek, P., editors, *INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*, pages 1194–1198. ISCA.

Yu, J., Lu, Q., Qin, Z., Yu, J., Li, Y., and Qin, Y. (2022). A multi-stage ensembled-learning approach for signal classification based on deep CNN and LGBM models. *Journal of Communications*, 17(1):30–38.

Zeng, Y., Mao, H., Peng, D., and Yi, Z. (2019a). Spectrogram based multi-task audio classification. *Multimedia Tools and Applications*, 78(3):3705–3722.

Zeng, Y., Mao, H., Peng, D., and Yi, Z. (2019b). Spectrogram based multi-task audio classification. *Multimedia Tools and Applications*, 78(3):3705–3722.

Zhang, T. and Zhang, X. (2022). Squeeze-and-excitation laplacian pyramid network with dual-polarization feature fusion for ship classification in SAR images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5.

Zhang, Y., Hong, D., McClement, D., Oladosu, O., Pridham, G., and Slaney, G. (2021). Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods*, 353:109098.

Zhu, B., Xu, K., Kong, Q., Wang, H., and Peng, Y. (2020). Audio tagging by cross filtering noisy labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2073–2083.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

Zissman, M. A. and Singer, E. (1994). Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. In *Proceedings of ICASSP '94: IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 305–308. IEEE Computer Society.

Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., and Chen, Y. (2015). Convolutional recurrent neural networks: Learning spatial dependencies for image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–26.