Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 395-398 recibido 04-12-2024 revisado 21-01-2025 aceptado 31-01-2025

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

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

Roberto Andrés Carofilis Vasco
Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León
Campus de Vegazana, s/n, 24007 León, España
andres.vasco@unileon.es

Abstract: This is the summary of the Ph.D. thesis conducted by Roberto Andrés Carofilis Vasco, under the supervision of Prof. Enrique Alegre Gutiérrez and Prof. Laura Fernández Robles at the University of León. The thesis defense took place in León, Spain, on December 20, 2023, in the presence of a committee formed by Dr. Luis Fernando D'Haro (Polytechnic University of Madrid, Spain), Dr. Kenneth P. Camilleri (University of Malta, Malta), and Dr. Victor González Castro (University of León, Spain). The thesis received international mention following a 3-month stay at the Idiap Research Institute in Switzerland, under the supervision of Dr. Petr Motlicek. The thesis was awarded with the outstanding *cum laude* distinction.
Keywords: Speech processing, language identification, accent identification, speaker identification.

Resumen: Este es el resumen de la tesis doctoral realizada por Roberto Andrés Carofilis Vasco, bajo la dirección del Prof. Enrique Alegre Gutiérrez y la Prof. Laura Fernández Robles en la Universidad de León. La defensa de la tesis se realizó en León, España, el 20 de diciembre de 2023 ante un tribunal compuesto por el Dr. Luis Fernando D'Haro (Universidad Politécnica de Madrid, España), el Dr. Kenneth P. Camilleri (Universidad de Malta, Malta), y el Dr. Victor González Castro (Universidad de León, España). La tesis obtuvo la mención internacional tras una estancia de 3 meses en el Idiap Research Institute, en Suiza, bajo la supervisión del Dr. Petr Motlicek. La tesis obtuvo la calificación de sobresaliente *cum laude*.
Palabras clave: Procesamiento del habla, identificación de idiomas, identificación de acentos, identificación de hablantes.

1 Introduction

Speech-processing models have gained increasing importance across various fields, including law enforcement and cybersecurity. These models play a crucial role in the fight against crimes like child exploitation and human trafficking, helping in suspect identification and providing 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 numerous challenges, a major one being the scarcity of data. Acquiring sufficient and relevant speech data poses an obstacle, as it makes it difficult to train models that show accuracy and robustness on specific tasks. Moreover, these models are often complex and require significant computational resources for both the training and inference phases, resulting in significant costs and time investments.

The thesis focuses on three speech-processing tasks: language identification, accent identification, and speaker identification. All three tasks are crucial in academia, industry, and cybersecurity, being useful in tasks such as victim and fugitive identification and tracking, crime prevention, and suspect segmentation. In addition, they have the potential to improve automatic speech recognition systems, addressing the challenges of creating robust systems that are resilient to the particularities of speech in different regions.

ISSN 1135-5948 DOI 10.26342/2025-74-28

In this thesis, we propose new techniques, models, and datasets to address the described speech processing tasks, which facilitate the creation of systems with state-of-the-art performance, and require a relatively low amount of data and computational resources to train. Motivated by our collaboration with the European project "Global Response Against Child Exploitation"(GRACE), we focus on the creation of applications that are useful for law enforcement agencies in their fight against cybercrime and child sexual exploitation.

Several contributions presented in this thesis will be used by Europol and the national law enforcement agencies of the European Union countries. The objective of the GRACE project is the creation of tools that allow the monitoring and generation of automatic alerts in cases of possible risk involving minors.

Among the proposals of this thesis are new systems capable of achieving competitive results even though they have been trained with a limited amount of data. In addition, it presents two new models capable of being trained with limited computational resources and at the same time achieving results superior to those of other state-of-the-art models.

We also include a new highly balanced dataset, and the experimental setup used in all the experiments carried out to allow reproducibility of the results and to make the results presented comparable with future tools.

2 Thesis Overview

This thesis is composed of 6 chapters, which are described below:

Chapter 1 presents the objectives, motivations, and introduces the contributions of this thesis.

Chapter 2 contains a detailed review of state-of-the-art approaches related to language identification, accent identification, and speaker identification tasks, and related work on the proposed contributions. We also mention the main limitations of the methods reviewed and possible 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" we present Grad-Transfer, a novel descriptor based on the concatenation of a flattened spectrogram and the dimensionality-reduced heatmaps generated with the Grad-CAM interpretability method (Selvaraju et al., 2020). This descriptor is capable of transferring knowledge extracted from a Convolutional Neural Network (CNN) specialized in accent identification, to be used as additional information in a Classical Machine Learning Algorithm (CMLA), to improve the results of the CMLA by enriching the data it receives as input.

We used Grad-Transfer for the classification of native English accents and compared it with the results achieved by CMLA and state-of-the-art deep learning models fed only by spectrograms. Grad-Transfer is especially useful in data-poor tasks, where CMLA may give better results than larger models.

The description of the pipeline, and the experimental results achieved, were published in the IEEE/ACM Transactions on Audio, Speech, and Language Processing journal (Carofilis et al., 2023a).

Chapter 4, entitled "MeWEHV: Mel and Wave Embeddings for Human Voice Tasks" presents a novel embedding enrichment procedure that combines the outputs of two concatenated models as independent branches of the same model. On the one hand, a branch with an embedding generation model fed by raw audio waves, called wave encoder, and, on the other hand, a branch with a CNN fed by MFCCs of the raw audios, called MFCC encoder.

We designed an architecture, named MeWEHV, capable of interacting with the two branches through a set of layers, including LSTM layers and attention mechanisms, combining the information extracted from both representations. MeWEHV was tested on the language identification, accent identification, and speaker identification tasks.

We empirically evaluated the hypothesis that there is a complementarity between the embeddings of the wave encoder, this being a non-imposed representation of the acoustic information, and the embeddings of the MFCC encoder, generated from MFCCs, this being an imposed representation.

We presented a new speaker identification dataset, named YouSpeakers204, which is highly balanced in terms of speaker accent and gender. We compared the MeWEHV model with six state-of-the-art models on the proposed tasks using nine datasets, including

YouSpeakers204.

Details of the MeWEHV architecture, dataset information, and experimental results were published in the IEEE Access journal (Carofilis et al., 2023b).

Chapter 5, entitled "Squeeze-and-excitation for embeddings weighting in speech classification tasks", presents the Squeeze-and-excitation for Embeddings Network (SaEENet), an update of the MeWEHV architecture. SaEENet is built using novel neural layers and several optimizations inspired by recent advances in other deep learning fields, such as the use of depthwise separable convolutions (Chollet, 2017), and squeeze-and-excitation blocks (Hu et al., 2020), initially proposed in the image processing field, and GRU layers (Cho et al., 2014), originally used in text processing.

In the SaEENet model, we introduce a novel implementation of squeeze-and-excitation block, which processes the stacked embeddings considering time as a dimension containing the target channels. Instead of weighting the relevance of 2D channels of a convolutional network, SaEENet weights each 1D embedding according to its relevance. This allows the next layer of the model to have the context of which embedding is more relevant, reducing the impact of embeddings generated from audio segments that do not contain speech or contain unnecessary information, and increasing the relevance of the segments that contain information of interest to the model.

We compared SaEENet with other state-of-the-art models, including MeWEHV, using three datasets, for the language identification, accent identification, and speaker identification tasks.

This chapter has been presented in an article detailing the work done and submitted to a journal.

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

3 Contributions

The main contributions of this thesis are presented below:

Grad-Transfer feature extractor. We introduced the new Grad-Transfer feature extractor to represent distinctive audio features that combine information from both the CNN-based class-discriminative localization technique Grad-CAM and spectrograms.

Novel accent classification approach. We proposed a new method for accent classification using Grad-Transfer, so that the method transfers knowledge from a CNN to a CMLA, achieving better results than other state-of-the-art models. This is the first time in literature to propose the use of a Grad-CAM-based method for knowledge transfer between machine learning models.

Benchmark setup for VCTK. We publicly present a setup for the Voice Cloning Toolkit (VCTK) dataset (Veaux et al., 2017) in the accent identification task, along with the results achieved by Grad-Transfer using that setup. With the aim that it can be used by researchers to test their models and compare the results with those of this work.

Multi-representation audio pipeline. We introduced a new pipeline to generate rich embeddings by merging multiple audio representations. This approach establishes a basis for improving large pre-trained models and increasing their performance without the need for retraining all their weights.

MeWEHV model architecture. Based on this pipeline we proposed the MeWEHV deep learning model architecture, which efficiently handles three speech classification tasks and achieves state-of-the-art performance on nine datasets. MeWEHV leverages the knowledge of frozen weights of pre-trained speech processing models and improves their performance by enriching the embeddings generated by them by adding information extracted from MFCCs, as a complementary representation. The MeWEHV architecture requires a relatively low number of trainable parameters, making it suitable for resource-constrained environments.

YouSpeakers204 dataset. We created a new dataset for speaker identification and accent identification, called YouSpeakers204, with 19607 audio clips and 204 speakers, which was created using public YouTube videos. The dataset is highly balanced according to the gender of the speakers and six accents: United States, Canada, Scotland, England, England, Ireland, and Australia.

Benchmarking Latin American Spanish Corpora. We used, for the first time in literature, the publicly available Latin American Spanish Corpora dataset (Guevara-Rukoz et al., 2020) in the accent identifica-

tion task, providing benchmark results with the systems we designed and an experimental setup made publicly available for reproducibility and future research.

SaEENet model architecture. We proposed SaEENet, a novel model architecture that achieves competitive results in speaker, language, and accent identification tasks. For the first time in the literature, we introduced the use of squeeze-and-excitation blocks to weight and filter compressed information in embeddings generated from audio clips.

Squeeze-and-excitation variants evaluation. We evaluated three variants of squeeze-and-excitation blocks and presented which variants work best for weighting embeddings of state-of-the-art models trained with self-supervised learning, and feature maps generated by a CNN.

State-of-the-art performance. We successfully outperformed the results of the MeWEHV model and other state-of-the-art models using the SaEENet architecture in the tasks of speaker identification, language identification, and accent identification. Among the other novelties of SaEENet are the use of depthwise separable convolution layers and GRU layers, reducing the number of trainable parameters.

Real-world application. We applied the models and techniques developed in this work to real-world scenarios, focusing specifically on extracting speaker information to identify offenders and victims. This work contributes to the efforts of the GRACE project to leverage machine learning techniques to combat child sexual exploitation.

Acknowledgements

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Framework Programme under the Global Response Against Child Exploitation (GRACE) Project under Grant 883341; in part by the Predoctoral Grant of the Junta de Castilla y León, under Grant EDU/875/2021; and in part by the framework agreement between the University of León and Spanish National Cybersecurity Institute (INCIBE) under Addendum 01.

References

Carofilis, A., E. Alegre, E. Fidalgo, y L. Fernández-Robles. 2023a. Improvement of accent classification models through grad-transfer from spectrograms and gradient-weighted class activation mapping. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 31:2859–2871.

Carofilis, A., L. Fernández-Robles, E. Alegre, y E. Fidalgo. 2023b. MeWEHV: Mel and wave embeddings for human voice tasks. *IEEE Access*, 11:80089–80104.

Cho, K., B. van Merrienboer, D. Bahdanau, y Y. Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. En D. Wu M. Carpuat X. Carreras, y E. M. Vecchi, editores, *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, páginas 103–111. Association for Computational Linguistics.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. En *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 1800–1807. IEEE Computer Society.

Guevara-Rukoz, A., I. Demirsahin, F. He, S. C. Chu, S. Sarin, K. Pipatsrisawat, A. Gutkin, A. Butryna, y O. Kjartansson. 2020. Crowdsourcing latin american spanish for low-resource text-to-speech. En N. Calzolari F. Béchet P. Blache K. Choukri C. Cieri T. Declerck S. Goggi H. Isahara B. Maegaard J. Mariani H. Mazo A. Moreno J. Odijk, y S. Piperidis, editores, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020*, páginas 6504–6513. European Language Resources Association.

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

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, y D. Batra. 2020. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 128(2):336–359.

Veaux, C., J. Yamagishi, K. MacDonald, y others. 2017. Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit.