IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 July 2023, accepted 30 July 2023, date of publication 1 August 2023, date of current version 4 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300973

RESEARCH ARTICLE

MeWEHV: Mel and Wave Embeddings for Human Voice Tasks

ANDRÉS CAROFILIS[1], LAURA FERNÁNDEZ-ROBLES[2], ENRIQUE ALEGRE[1], AND EDUARDO FIDALGO[1]

[1]Department of Electrical, Systems, and Automation Engineering, School of Industrial, Computer and Aerospace Engineering, Universidad de León, Campus de Vegazana, 24007 León, Spain
[2]Department of Mechanical, Computer, and Aerospace Engineering, Universidad de León, Campus de Vegazana, 24007 León, Spain

Corresponding author: Andrés Carofilis (andres.vasco@unileon.es)

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; and in part by the ''Ayudas para financiar la contratación predoctoral de personal investigador'' grant of the Government of Castilla y León, Spain, under Grant EDU/875/2021.

ABSTRACT A recent trend in speech processing is the use of embeddings created through machine learning models trained on a specific task with large datasets. By leveraging the knowledge already acquired, these models can be reused in new tasks where the amount of available data is small. This paper proposes a pipeline to create a new model, called Mel and Wave Embeddings for Human Voice Tasks (MeWEHV), capable of generating robust embeddings for speech processing. MeWEHV combines the embeddings generated by a pre-trained raw audio waveform encoder model, and deep features extracted from Mel Frequency Cepstral Coefficients (MFCCs) using Convolutional Neural Networks (CNNs). We evaluate the performance of MeWEHV on three tasks: speaker, language, and accent identification. For the first one, we use the VoxCeleb1, and VBHIR datasets and present YouSpeakers204, a new and publicly available dataset for English speaker identification that contains 19607 audio clips from 204 persons speaking in six different accents, allowing other researchers to work with a very balanced dataset, and to create new models that are robust to multiple accents. For evaluating the language identification task, we use the VoxForge, Common Language, and the LRE17 datasets. Finally, for accent identification, we use the Latin American Spanish Corpora (LASC), Common Voice, and the NISP datasets. Our approach allows a significant increase in the performance of state-of-the-art embedding generation models on all the tested datasets, with a low additional computational cost.

INDEX TERMS Embeddings, HuBERT, speech classification, WavLM, XLSR-Wav2Vec2, YouSpeakers204.

I. INTRODUCTION

Speech processing refers to analyzing human speech through voice audio signals. Some of the most important problems in this field, which are tackled in this paper, are language identification, accent identification, and speaker identification [1], [2], [3].

On the one hand, language identification identifies the spoken language present in an audio file, and accent identification determines a person's region of origin based on the characteristic way and tone of the language used. We consider that the existence of very similar languages or accents, usually languages or accents with a common origin, poses a challenge for both tasks and requires the use of powerful machine learning models. In some speech processing tasks, there are useful datasets publicly available [4], [5], [6], [7], [8], [9], but the comparison of results on such datasets becomes difficult due to the lack of a common predefined experimental setup for training the models and the lack of previous research results to compare with, such as the case of accent detection in Spanish [10].

On the other hand, speaker identification consists of recognizing the identity of a person given an audio file with a person's voice. A problem in this field is that there is a lack of

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

well-balanced English datasets, both in the number of audios per speaker and the number of speakers per accent [4], [7], [8], [9], [11], [12]. This can lead to the creation of models that might not identify accents in real-world data effectively. The availability of a dataset with these features would allow the creation of more effective models on real-world problems and facilitate the integration and evaluation of multiple tasks, such as speaker identification and accent identification, simultaneously.

Most of the research focused on the three aforementioned tasks addresses them individually, and the proposed systems are usually evaluated for just one task [1], [2], [3], [13], [14], [15]. However, some research addresses several of these problems based on the same pre-training model [16], [17], [18], [19].

A machine learning architecture capable of performing well on multiple speech processing tasks can use the knowledge acquired during training, with a large amount of data, and exploit it in new and diverse tasks. In this way, the model generated for a new task does not need to start from scratch, thus requiring less training data. By freezing the trained layers, fewer parameters would need to be trained, with a consequent reduction in the computational power required [16], [18], [19].

Various techniques exist for reusing these models on specific tasks other than those for which they were initially trained. This field of research is known as *transfer learning* [20]. Some models address the transfer learning problem by creating deep representations [17], [18], [19], [21], [22], also called *embeddings*. An embedding represents a position in an abstract multidimensional space that 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 [23]. The embeddings have been used in multiple domains, such as text, image, and speech processing, and can feed multiple systems for an individual task in each one [24], [25].

For the speech processing domain, there are models that address speech classification for one or more tasks using embeddings. For example, WavLM [19], presented as a universal speech encoder, was tested in tasks such as speaker identification, and speech to text, among others. There are also models that, although they were developed for a specific task, are also capable of creating embeddings, so they can be reused in new tasks. This is the case of HuBERT [18] and Wav2Vec2 [17], which are focused on English speech-to-text conversion, and XLSR-Wav2Vec2 [21], which is based on Wav2Vec2 but adds the possibility to work with multiple languages.

The embedding generation models mentioned above were trained with thousands of hours of audios recorded in multiple environments, resulting in models that can generate robust embeddings against background noise and different environmental conditions. This paper takes into account the advances achieved by this class of models and leverages them for speaker identification, accent identification, and language identification, by means of transfer learning.

We develop and present a novel embedding enrichment procedure, which combines the outputs of two models. On the one hand, an embedding generation model from raw audios, which we refer to, in a general way, as *wave encoder*. On the other hand, the outputs of a neural network (NN) fed by the Mel Frequency Cepstral Coefficients (MFCCs) [26] of the raw audios, which have among its advantages the capability of error reduction and robustness to noise [27]. The main feature of MFCC is that it focuses on extracting relevant audio components to identify speech features, discarding, by filtering, other features such as background noise, pitch, loudness, and emotion, among others. Therefore, we call the NN an MFCC encoder.

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. The information contained in the raw audio may contain relevant information that may have been filtered out in the MFCC, and the MFCC provide the machine learning model with information on the most relevant parts of an audio, on which it should focus.

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 Soft Attention layers.

The LSTM layer is a type of recurrent neural network layer that effectively captures long-term dependencies in sequential data by incorporating a memory cell, allowing it to retain and utilize information over extended sequences. A Soft Attention layer dynamically focuses on different parts of the input sequence, assigning varying levels of importance to each element. By combining an LSTM layer with a Soft Attention layer, the model gains the ability to capture long-term dependencies while selectively attending to crucial elements.

With the proposed architecture we managed to overcome the results obtained by other state-of-the-art embedding generation models, at the same time requiring only a small number of trainable parameters. Fig. 1 shows a basic scheme of the proposed architecture.

This paper provides the following main contributions:

- Proposal of the MeWEHV (Mel and Wave Embeddings for Human Voice Tasks) model architecture, which efficiently handles multiple speech classification tasks and achieves state-of-the-art performance. It leverages frozen weights from pre-trained models and requires a relatively low number of trainable parameters, making it suitable for resource-limited environments.
- Introduction of a pipeline for generating rich embeddings by merging multiple audio representations. This approach establishes a basis for improving large pre-trained models and enhancing their performance.

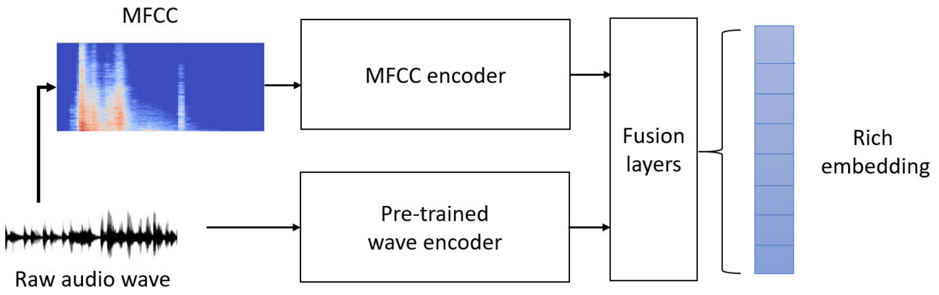

FIGURE 1. A basic representation of the proposed architecture. It merges two types of representations and generates rich embeddings.

- Creation and presentation of the YouSpeakers204 dataset, a balanced speaker identification dataset with diverse speaker accents and gender, extracted from publicly available YouTube videos.
- Novel use of the Latin American Spanish Corpora for accent identification, providing baseline results and an experimental setup for future research in this domain.
- Application of the research to real-world scenarios, specifically focusing on speaker information extraction for identifying offenders and victims. This work contributes to the GRACE project's efforts in leveraging machine learning techniques to combat child sexual exploitation.

The remaining part of the paper is organized as follows: In Section II, a review of the state of the art in the field of speech processing is presented. Then, in Section III, the information of the new YouSpeakers204 dataset is introduced. In Section V, the proposed architecture is described. In Section VI, the rest of the datasets and the experimental setup are detailed. In Section VII, the results obtained with each of the datasets are described. Finally, the discussion and the conclusions obtained are given in Sections VIII and IX, respectively.

II. STATE OF THE ART

In the field of speech classification, multiple solutions have been developed for a single task. Reference [13] 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.

Reference [3] 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 [28], TAL_ASR, and a combination of the datasets THCHS-30 [29] and TIMIT [4], 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-SVM and x-vectors.

In speaker identification, [1] introduced CASA-GMM-CNN model, in which they seek to clean a noisy audio through a Computational Auditory Scene Analysis (CASA), then make a classification of emotions through a GMM-CNN, and the output of both components feed another GMM-CNN in charge of identifying the speaker. They tested their approach on SUSAS [30], Arabic Emirati Speech Database [31], RAVDESS [32], and Fluent Speech Commands [33] datasets, achieving an improvement in accuracy of up to 59.37% with respect to other state-of-the-art works.

Reference [14] 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.

In accent identification, [2] presented AISpeech-SJTU, an accent identification system that is powered by Phone Posteriorgrams and data augmented by text-to-speech synthesis systems. They evaluated their proposal in the Interspeech-2020 Accented English Speech Recognition Challenge [34], achieving an average accuracy of 83.63%, the highest score of the challenge.

Transfer learning and domain transfer have been extensively studied in machine learning [35]. Recent research related to transfer learning in audio processing has mainly focused on methods for learning deep representations, also known as embeddings [16]. These embeddings are generated to store relevant information of an audio wave, through its representation in a latent space, to be later used in the learning of a new specific task.

Reference [15] 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.

One powerful model focused on the generation of embeddings is TRILL [16], which was trained with a subset of the AudioSet dataset [36], and subsequently evaluated in different domains by applying 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 [11], 94.1% for language identification on the VoxForge dataset (5.7% improvement) [5], 91.2% for command identification on the Speech Commands dataset [37] (0.1% improvement), among others.

Other embedding generation models are the Wav2Vec2 [17] model, which focused on English speech-to-text conversion, and XLSR-Wav2Vec2 [21] 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 [12], CommonVoice [6], and BABEL [9] datasets were used. The XLSR-Wav2Vec2 model is fed by the raw audio waves and 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.

Both models are based on the transformer architecture and are trained with self-supervised learning tasks using large audio datasets.

The Wav2Vec2 model and the XLSR-Wav2Vec2 model achieve outstanding performances, outperforming smaller state-of-the-art models, and have shown to effectively capture and model long-term dependencies in sequential speech data.

Reference [18] 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 [7] and Libri-Light [8] datasets, achieving WER improvement of up to 19%.

In the experimental results of [18], HuBERT showed better results than Wav2Vec2 in low resource setups, although the size of both models is similar (318.42M in the case of Wav2Vec2-large, and 314.65 in the case of HuBERT-large). However, unlike XLSR-Wav2Vec2, HuBERT does not present a specific configuration for multiple languages.

Reference [19] 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 [38] achieving an overall score 3.16% higher than HuBERT and 4.95% higher than Wav2Vec2.

Although the large version of the WavLM model is relatively large (317.66 M parameters) and requires large computational resources for training, the authors proposed among their future lines of research the enlargement of their models as a method to increase the performance achieved by them.

For the correct democratization of the most powerful deep learning models, we consider that it is crucial to explore alternative methods for enhancing their performance without solely relying on increasing model size. This approach should involve the utilization of a limited number of trainable parameters, enabling researchers to adapt these large models to their needs, even when large computational cluster may not be available.

Apart from the models focused on speech processing, there are also models for general audio processing, such as the PANN model. The PANN model [39] 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 [40] and the MSoS [41] datasets, respectively, surpassing the state-of-the-art results.

For acoustic scene classification, PANN was evaluated on the datasets DCASE-2019 [42] and DCASE-2018 [43], 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 [44], 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.

Approaches based on embedding generation have demonstrated competitive performance in multiple audio processing tasks using transfer learning. 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 been focused significantly on the use of a single representation of the audio. Among the most common representations are the use of spectrograms [45], [46], [47], and MFCCs [48], [49], which can be competitive depending on the task and the dataset used, and, in general, both can obtain similar results [50].

Different representations and features extracted from an audio can be used at the same time to feed a model. One example is FuzzyGCP [51], 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 [52], IIT Madras [53], VoxForge, and MaSS [54], 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 [55], i-vector [56] and x-vector [57].

The combination of representations makes possible to extract complementary information from the original audios, in a format easily processed by a machine learning model. This allows these models to achieve better results than those obtained by being fed by a single representation.

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. However, in our paper, we use, among the evaluation datasets, a subset of the VoxForge dataset, which was created based on the general data provided in the paper.

Another model based on the combination of representations is the model proposed by [58], in which they combined three types of audio representations, which 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 93.3% in the acoustic scene classification 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 into model outputs has a limitation given that the information that can be shared in this way is limited, compared to the information that could be obtained if deep representation were connected.

Reference [59] proposed a novel architecture fed by three types of representations, these representations fed two consecutive NN. 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 to be especially valuable in cases where the training data are not properly filtered, which can affect the performance of models that, in particular, are trained with relatively small datasets. In our case, we take as reference embedding generation models that have been trained with large amounts of data makes the resulting pre-trained model highly robust and resistant to possible errors in the training set.

In our paper, we address the combination of representations from a novel approach, taking advantage of the proven capabilities of embedding generation models and improving those capabilities by means of a new architecture that complements the information generated by them. Unlike other approaches, our method requires a single end-to-end robust model. Considering the described works, it can be noted that one of the major limitations in embedding generation models is the fact that retraining these models is computationally expensive and requires a large amount of data. For this reason, the embeddings generated by these models are usually reused through the application of transfer learning, enabling the utilization of the knowledge already acquired during their initial training process to address new tasks.

We propose the MeWEHV 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. Using the combination of both types of embeddings we are able to surpass the state-of-the-art results in multiple speech processing tasks, taking advantage of the benefits of embedding generation models and combination of representations.

The MeWEHV architecture allows to improve the results obtained with transfer learning, through the enrichment of embeddings. In this way, competitive results can be achieved, without the need to retrain the complete models.

In addition, MeWEHV opens the door to future research that seeks to adapt large speech processing models by improving the use of available speech data, without requiring a significant increase in the number of parameters.

III. YouSpeakers204 DATASET

We introduce a new dataset for speaker identification, called YouTube Speakers 204 (YouSpeakers204), which contains 19607 audio clips of 204 speakers with 6 different accents extracted from YouTube videos. We selected YouTube channels in which the information of the country of origin and gender of the speaker was stated and looked for native English speakers with a wide range of ages, ethnicities, and professions. Their respective YouTube channels contain different topics. All speakers are native English speakers, grouped by region of origin. The 6 labeled regions are the United States, Canada, Scotland, England, Ireland, and Australia. The dataset is gender-balanced, with 50% male speakers and 50% female speakers. The audios included in this dataset come from varied recording environments, including indoor studios, outdoor recordings, professional recordings, and recordings with background noise.

The presented dataset is intended to facilitate research in the field of automatic speaker classification, and can also be used in related studies combining speaker identification with accent identification.

YouSpeakers204 contains data recorded in noisy environments under a wide variety of real conditions, such as the recording microphone used, background noise, recording environment, audio volume, speaker gender, and accent, which makes the dataset a challenge to test the robustness of new machine learning models.

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

TABLE 1. YouSpeakers204 dataset statistics.

# of speakers	204
# of male speakers	102
# of female speakers	102
# of accents	6
# of videos	1055
# of minutes	2026
# of audio clips	19607
Avg. # of videos per speaker	5.17
Avg. # of clips per speaker	96.11
Avg. length of clips	6.19 seconds

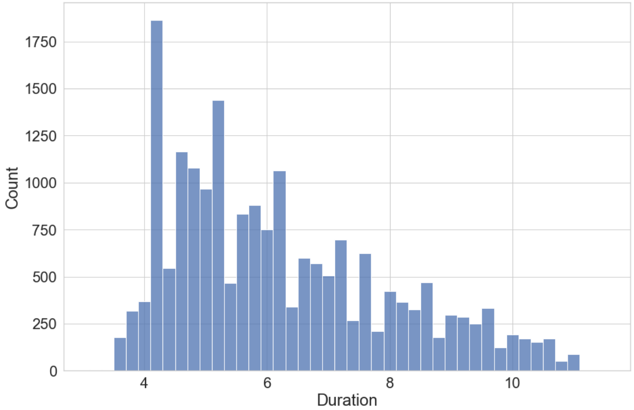

FIGURE 2. Distribution of clip duration in seconds in the YouSpeakers204 dataset. The audios are between 3.5 seconds and 12 seconds in length.

The general statistics of the dataset are presented in Table 1, while Fig. 2 includes the length distribution of the audio clips.

For the creation of YouSpeakers204, we defined a procedure, which consists of the following stages: listing candidate speakers, selecting and downloading videos, and audio processing.

A. LISTING CANDIDATE SPEAKERS

The list of speakers was extracted from the Socialblade[4] database by selecting the most famous YouTubers for each region among the 6 accent classes. In the case of Scotland and England, it was necessary to perform a manual search within YouTube to find people who, in their public information, claim to be from those regions, because Socialblade divides YouTubers by countries and not by regions. Subsequently, a verification of the place of birth of each speaker was carried out, by a search in Wikipedia.[5] The collected list contains a total of 204 speakers, 34 per accent of which 17 are men and 17 are women ($(17 + 17) \times 6 = 204$). All speakers have been assigned a unique identifier (id) and their real identity is not provided in the dataset.

B. SELECTING AND DOWNLOADING VIDEOS

To create a diverse dataset, 19607 clips were extracted from a large number of videos (1055 videos in total). In this way, models created using YouSpeakers204 dataset can be robust to the different environments and contexts in which the audios had been recorded. Each speaker has an average of 5.17 videos from which their clips were extracted, an average of 96.11 clips, and each clip has an average length of 6.19 seconds.

After selecting the videos, the entire video was downloaded and the audio clips were extracted.

C. AUDIO PROCESSING

The complete audios were processed manually, by a team of taggers, separating the original audios into segments of short duration and storing the resulting audios together with their respective information. The process consisted of defining a decibel split threshold for each audio, all generated sections with a decibel level below the threshold are considered silences (see Fig. 3). All segments are obtained by extracting the sections that are between every two contiguous silences. Due to the particularities of each audio, it is not possible to define a unique threshold that allows to label the silences of all the audios in a correct way. Therefore, applying a visual analysis of the audio waves the threshold of each file was defined manually.

The segmentation of the audios generates multiple clips of variable size, of which only the clips with a duration between 3.5 and 12 seconds are kept. A manual check of the content of each clip is then performed to discard all clips containing voices other than the target YouTuber, and clips containing no voice.

Finally, the resulting clips are renamed and stored, the names of the files contain an anonymized speaker id, anonymized id of the video from which each clip originates, the gender of the speaker, and the region of the speaker, which represents the accent.

IV. THEORETICAL FRAMEWORK

A. PRETRAINED EMBEDDINGS GENERATION MODELS

The embeddings generation models used in this paper are large deep learning models developed for learning self-supervised representations of speech data, known as embeddings. These models are trained on large unlabeled datasets to learn to generate embeddings, without the need for explicit phonetic or linguistic annotations.

In this paper, these pre-trained models are used in the wave encoder block of the MeWEHV architecture, and the embeddings generated by them are enriched using the pipeline of the proposed architecture. Thus, multiple models based on the MeWEHV architecture were evaluated, one for each wave encoder analyzed.

The layers of the embeddings generation models can be grouped into two categories, according to their functionality. On the one hand, the encoder layers take as input a raw

[4] https://socialblade.com
[5] https://www.wikipedia.org

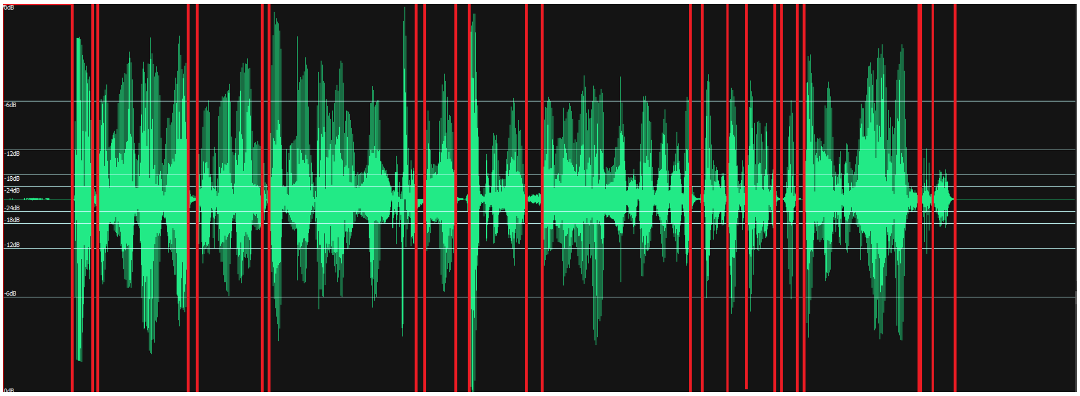

FIGURE 3. Example of an audio cut-off in "silence" regions, i.e. regions where the decibels are below a certain threshold. The audio waveform is shown in green and the cut-off regions are in red.

audio, divide it into fixed size sections (usually 20-30 ms) $S = S_1, S_2, \ldots, S_N$ and generate a set of embeddings $E = E_1, E_2, \ldots, E_N$, one per section. Each embedding E_i, where $1 \leq i \leq N$ represents the relevant features of a section S_i on the training process for its target task. On the other hand, for the model training process, another set of layers called *decoder* was used, which takes the embeddings and processes them to generate the expected output according to the assigned task. In our architecture, the wave encoder block is composed of the encoder layers of each of these models.

The embeddings generated by the wave encoder represent a position in the latent space, where audios with similar characteristics are represented spatially close together, and audios with different characteristics are represented far apart.

B. MFCCs

The MFCCs are coefficients based on the human audible frequency range, represented by the Mel scale, which is a linear scale below 1000 Hz and logarithmic above 1 kHz [26]. The Mel scale is a scale that relates the perceived frequency of a tone to the actual measured frequency. It modifies a frequency to approximate what the human ear can hear and is often used to extract features of an audio signal that are relevant to identifying its content, making it useful in tasks such as speech representation.

The MFCCs reduce the relevance of information that may have a minor contribution to speech-processing tasks, which may add noise to the model and reduce its accuracy.

C. CENTER LOSS FUNCTION

The center loss [60] is a loss function that enhances the discriminative power of the learned features by minimizing the distances between the features and their respective class centers. It introduces class centers and pulls the feature vectors toward these centers during training, promoting compact clustering of features belonging to the same class, through the following equation:

$$L_c = \frac{1}{2} \sum_{i=1}^{N} \left\| \mathbf{x}_i - \mathbf{c}_{y_i} \right\|_2^2 \tag{1}$$

where N is the number of samples, $\mathbf{x}_i$ is the feature vector of the i-th sample, y_i is its corresponding class label, and $\mathbf{c}_{y_i}$ is the center of the class y_i.

V. ARCHITECTURE: MEL AND WAVE EMBEDDINGS FOR HUMAN VOICE TASKS

Fig. 4 depicts a summary of the Mel and Wave Embeddings for Human Voice Tasks (MeWEHV) architecture.

The MeWEHV architecture is fed by two inputs, on the one hand, the audio signal is treated as a one-dimensional vector, and on the other hand, the MFCCs [26] are extracted from the same audio signal.

A. WAVE ENCODER BRANCH

The two inputs of the MeWEHV model are processed independently by two branches. On the one hand, the raw audio waveform feeds the encoder layers of a B_2 wave encoder model. The B_2 wave encoder is a block of the architecture that is composed of a pre-trained embedding generation model. In this paper, multiple models were tested as the wave encoder block.

The models used as wave encoders are XLSR-Wav2Vec2, HuBERT in its base and large versions, and WavLM in its base and large versions. In future research, these models could be replaced by others.

The first branch of our model will generate multiple embeddings E_i, one for each audio section S_i, which will summarize the features that the wave encoder considers relevant. We perform transfer learning by feeding the generated embeddings to new layers connected to the encoder outputs.

We connected the embeddings of the generated sections to an LSTM layer, L_2, and a Soft Attention layer, A_2 that will be able to model their temporal information and generate a single embedding for the complete audio signal. The blocks that compose this branch are $B_2 + L_2 + A_2$.

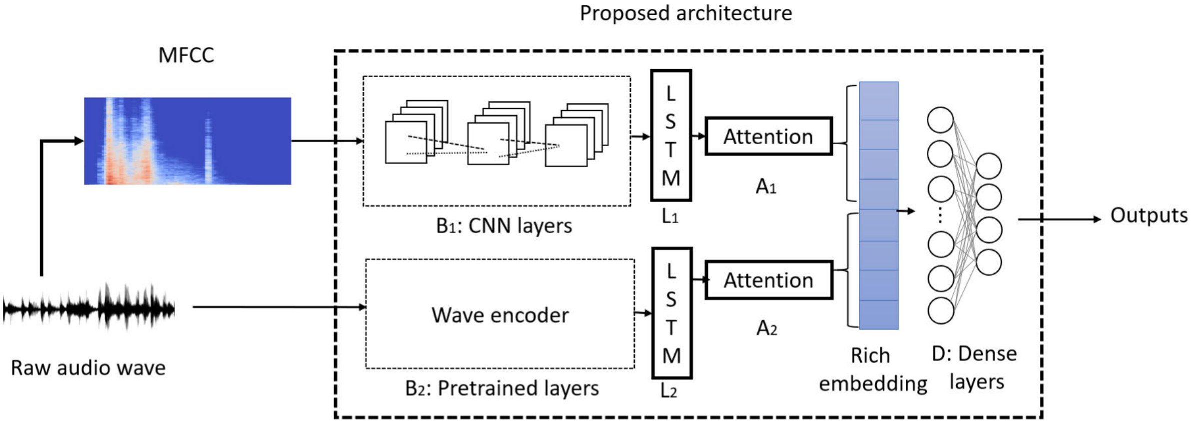

FIGURE 4. The architecture of the MeWEHV model proposed in this paper. The model is fed by two inputs, on the one hand, the raw audio waveform that feeds the wave encoder branch, which contains the encoder layers of a pre-trained embedding generation model from raw audios. On the other hand, the MFCC coefficients extracted from the same raw audio feed the MFCC branch which contains a set of convolutional layers. Both branches of the model pass through an LSTM layer and a Soft Attention layer independently. These layers are in charge of modelling the temporal features and generating, each one, a unique embedding from each input audio. Both embeddings are concatenated and subsequently feed a series of dense layers that are responsible for the classification of the model.

B. MFCC BRANCH

The second branch of the proposed model is fed by the information contained in the MFCCs of the original audios and generates a new embedding. In this way, we enrich the embeddings generated by the first branch of the model.

The use of MFCCs allows us to analyze the most important information of an audio and complement the information that the wave encoder models may have missed during their analysis.

We use the B_1 block, composed of three concatenated layer sets to process the MFCCs, each set consists of a 1D convolutional layer, a batch normalization layer, and a ReLU activation function. At the same time, the output of the last block is connected to an LSTM layer, L_1, and a Soft Attention layer, A_1, in the same way as the first branch of the model. The output of this layer block is a new embedding.The blocks that compose this branch are $B_1 + L_1 + A_1$.

C. RICH EMBEDDINGS

The embedding generated by the first branch and the embedding generated by the second branch are concatenated and generate a new embedding which size is the sum of the embedding sizes of both branches. The generated embedding is optimized through the center loss function.

Thus, the result of the concatenation of both embeddings is a rich embedding in a new latent space.

Finally, the rich embedding feeds a block D of two fully connected neuron layers, with intermediate ReLU activation and a dropout function, which is responsible for generating the output of the model that classifies the input according to the assigned task.

The center loss score is combined with the classification loss of block D, in this case, the negative log-likelihood [61]. In particular, the risk function used during training is the sum of the center loss and the classification loss. The objective is to minimize both the classification loss and the distance between the learned features and their class centers.

The proposed MeWEHV architecture complements the information of the pre-trained wave encoder model with the information extracted using the MFCCs to generate a more powerful and flexible model.

VI. EXPERIMENTS

A. DATASETS

The datasets used allow us to evaluate the performance of our model in the tasks of language identification, accent identification, and speaker identification. These datasets are diverse in terms of the number of speakers, nationalities, gender, and environments in which they were recorded, which allows us to evaluate the correct functionality of a MeWEHV model with complex data.

The lists of audios used in the training, validation, and test partitions are publicly available,[6] so that future research can make a fair comparison of results.

1) VoxForge

VoxForge is a dataset composed of the voices and transcriptions of a large number of speakers, originally intended for speech-to-text conversion. It comprehends a large number of languages, which makes it also useful for language identification.

We use this dataset to compare the proposed architecture in language identification. The used subset is based on the FuzzyGCP paper presented by [51].

This subset contains 5 languages: French, German, Italian, Portuguese, and Spanish. All the speakers of each language were divided into a proportion of 70% for training, 10% for test, and 20% for validation. Subsequently, considering only the audios with the selected speakers, 1400 audio clips

[6] Partitions of the datasets used are publicly available at: https://bit.ly/3ydSEAt

were randomly chosen for the training set, 200 for the test set, and 400 for the validation set, per language. This results in a training, test, and validation sets with 7000, 1000, and 2000 audios, respectively.

The generated partitions have no speaker contamination, i.e., the speakers present in one set are not present in the other sets, which assures that the model learns to recognize the languages and not the voices of the speakers.

2) COMMON LANGUAGE

The Common Language dataset [62] has a set of audios selected from the Common Voice dataset [63], which contains audios provided by volunteers. Common Language contains 45 languages, 272360 audios, and 13808 speakers.

The dataset was used for language identification and we used the training, validation, and test partitions provided by the authors: 177552, 47104, and 47704 audios for training, validation and test sets, respectively.

3) LRE17

The NIST 2017 Language Recognition (LRE17) [64] is a dataset created for the language identification task, with approximately 2100 hours of audio in 14 languages. The dataset comprises a training set with 15904 audios, and a predefined test set with 25451 audios.

For the training of the models used in this paper, the original training set was randomly divided into 90% for training and 10% for validation. Subsequently, due to the long duration of the original audios, for both subsets the audio segments containing speech were extracted, and in order to eliminate examples with little relevant information, segments less than 1.5 seconds long and segments without speech were discarded. Each generated segment was treated as a separate example during the training and validation phase.

The resulting subsets consisted of 616324 training and 67451 validation examples. In the test set, the original 25451 audios, without separation into speech segments, were used.

4) LATIN AMERICAN SPANISH CORPORA

Latin American Spanish Corpora [10] was originally proposed for the speech-to-text conversion task. However, thanks to being a highly balanced dataset, both by gender and by accents, it can be used in accent identification. In this paper, to the best of our knowledge, this is the first time that this dataset is being used for for accent identification.

The dataset comprises 37.79 hours of 6 Latin American accents: Argentinian, Chilean, Colombian, Peruvian, Puerto Rican, and Venezuelan. We divided the speakers of the dataset into training (70%), validation (15%), and test sets (15%), and use all the audios of each speaker.

5) COMMON VOICE

We used the Common Voice dataset for the task of accent identification. We worked with an English subset containing audios of five accents: American, British, Indian, Canadian, and Australian.

We used a subset with 10000 audios per accent, which were divided into training (70%), validation (15%), and test sets (15%), resulting in a training set with 35000 audios, a validation set with 7500 audios, and a test set with 7500 audios.

6) NISP

The NISP [65] dataset is a speaker profiling dataset containing information such as height, age and accent of 345 speakers. The speakers have 5 mother tongues (Hindi, Kannada, Malayalam, Telugu, and Tamil), and among the audios in the dataset there are English recordings of these speakers.

This dataset was used for the accent identification task, and, for the experimental phase of this paper we consider each of the mother tongues as accents, and, therefore, as the true labels to be predicted by the evaluated models.

For the partitioning of the dataset, the 345 speakers were divided randomly and stratified in order to maintain the same proportions in each accent. The portions used are: 70% for training, 15% for validation, and 15% for testing. Subsequently, all the audios belonging to each speaker were used as part of their corresponding partition. Thus, the final training set contains 10247 audios, the validation set 2246 audios, and the test set contains 2201 audios.

7) VoxCeleb1

For the speaker identification task, another of the datasets we chose is VoxCeleb1 [66], which is composed of 153516 audios samples from 1251 different speakers. The audios of the dataset were extracted from public videos of celebrities on YouTube.

For the VoxCeleb1 dataset, the same partition proposed by the creators [7] was used, which contains 138361 audios in the training set, 6904 audios in the validation set, and 8251 audios in the test set.

8) YouSpeakers204

The YouSpeakers204 dataset is one of the contributions of this paper and its information can be found in Section III.

The entire YouSpeakers204 dataset was used for speaker identification. The audios of each speaker were divided into a proportion of 70% for training, 15% for validation, and 15% for test, and all the audios available in the dataset were used. This resulted in a training set of 13728 audios, 2942 audios in the validation set, and 2942 audios in the test set.

9) VBHIR

A Dataset for Voice-Based Human Identity Recognition, which we call VBHIR [67] is a dataset containing 3000 audios from 150 English speakers of Middle Eastern descent. Half of the audios in the dataset contain 10 recordings per speaker, in which they read the same text, while the other half contains 10 recordings per speaker, where they read

[7] https://www.robots.ox.ac.uk/∼vgg/data/voxceleb/meta/iden_split.txt

a different text in each audio. In the second case, the texts read by the speakers do not repeat with the other speakers.

In the present paper, in order to evaluate the robustness of the proposed models, we considered only the recordings where a different text is read, resulting in a dataset of 1500 audios.

The dataset was divided into training, validation, and test subsets, with a proportion of 70%, 15%, and 15%, respectively. Resulting in a training set with 1050 audios, a validation set with 225 audios, and a test set with 225 audios.

B. EXPERIMENTAL SETUP

For the experimentation, in all the datasets and tasks, audios with a sample rate of 16000 samples per second were used, converted into 8-second clips. Those with a shorter duration were repeated as many times as necessary until reaching 8 seconds, and those with longer duration were trimmed and only the first 8 seconds were worked on.

In the creation of the MFCCs, 128 MFCC coefficients were defined as a parameter to be used, which we consider it provides the MFCC with a high level of spectral detail, which allows the models to perform tasks requiring such detail.

The specific parameters of the MeWEHV model used in our experimentation can be seen in Table 2, which were empirically selected.

We established multiple baseline models, on which the MeWEHV architecture is applied, and took them as a reference to compare the performance of our proposal. Among the architectures used as baseline we include the CNNMFCC architecture, which is composed of the same layers contained in the MFCC branch of the MeWEHV model, described in Section V.

In addition, we include each of the embeddings generation models used as wave encoder in the MeWEHV architecture as standalone models. These models are composed of the same layers as the wave encoder branch described in Section V. In Section VII, all models generated from the wave encoder branch structure are named identically to their corresponding embedding generation model.

The architectures based on the wave encoder branches are presented in Table 3. These architectures are composed by the encoder layers of the pre-trained embedding generation model and the classification layers of the first branch of the MeWEHV architecture. The layer blocks B_2, L_2, and A_2 presented in Table 3 have the same experimental configuration and number of parameters as those described in Table 2, with the same name. The only difference appears in block D_2, which uses 128 neurons instead of 256 of block D presented in Table 2, resulting in a model with $625,542$ trainable parameters.

Overall, the wave encoder branch-based architectures are composed of the $B_2 + L_2 + A_2 + D_2$ layers.

The structure of the CNNMFCC models, based on the MFCC branch, is presented in Table 4. The B_1, L_1, and A_1 blocks described have the same experimental configuration and number of parameters as the blocks with the same names in Table 2. As with the wave encoder branch-based architectures, in this case, the fully connected layers of block D_1 has 128 neurons, unlike the 256 of block D in Table 2, resulting in a model with $355,654$ trainable parameters.

Overall, the CNNMFCC architecture comprises the blocks $B_1 + L_1 + A_1 + D_1$.

For both the wave encoder branch-based architectures and the CNNMFCC architecture, we use the same experimental setup as presented in each respective branch of the MeWEHV architecture. This approach ensures a fair comparison of results across all models.

VII. RESULTS

The results of the experiments performed can be seen in Table 5. Since the evaluated datasets are balanced in their respective classes, we use accuracy as the metric in the experimentation.

The number of parameters mentioned in Table 5 for the different versions of XLSR-Wav2Vec2, HuBERT and WavLM include the 0.62M trainable parameters of the LSTM, Soft Attention, and classification layers added, which have as input the embeddings generated by each of the mentioned models.

As it can be seen, the implemented MeWEHV models improve the results with respect to all the embedding generation models used as wave encoders and on which our proposal was implemented. It is worth noting that the MeWEHV models have only 0.68M more parameters than their corresponding baseline models, which represents, in the case of the XLSR-Wav2Vec2 model, the baseline model with the highest number of parameters, an increase of only 0.21% of parameters.

On the VoxForge and Common Language datasets, the best language identification model is MeWEHV, using WavLM-large as wave encoder, achieving accuracies of 97.06% and 72.53%, respectively, which represents an improvement of 0.73% and 14.47% with respect to WavLM large, with the improvement achieved with Common Language being the largest among the models tested in this dataset with respect to its baseline model. The largest improvement with VoxForge was achieved with the XLSR-Wav2Vec2-based MeWEHV model with respect to the XLSR-Wav2Vec2 model and represents an increase of 14.86%.

The MeWEHV model, using WavLM-large as wave encoder, also achieved the best result with the LRE17 dataset, yielding an accuracy of 41.05%, which represents a 12.49% of improvement in language identification compared to the baseline WavLM-large. Furthermore, the largest leap in performance on the LRE17 dataset was achieved with the MeWEHV-XLSR-Wav2Vec2 model, obtaining an improvement in accuracy of 46.68%, compared to the XLSR-Wav2Vec2 model.

In addition to the mentioned results, we can add as a baseline the result achieved by the FuzzyGCP model in language identification with the VoxForge dataset, obtaining

TABLE 2. Details of the MeWEHV model used in the experimentation, assuming a task with six possible output classes. *(T)* represents that the input of a given layer has been transposed. With each extra output class, the number of parameters increases by 256, being the number of connections that the new output neuron would have with the penultimate layer. The number of wave encoder parameters depends on the model chosen for that block, and the size of the generated embeddings of 1024 was assumed as the wave encoder output.

Layer blocks (index. id: type: depth)	Layers	Input shape	Output shape	Param #	Kernel shape
1. B_1: CNN: 1	Conv1d	[128, 641]	[128, 319]	82,048	[128, 128, 5]
	BatchNorm1d	[128, 319]	[128, 319]	256	[128]
	ReLU	[128, 319]	[128, 319]		
	Conv1d	[128, 319]	[128, 316]	65,664	[128, 128, 4]
	BatchNorm1d	[128, 316]	[128, 316]	256	[128]
	ReLU	[128, 316]	[128, 316]		
	Conv1d	[128, 316]	[128, 313]	65,664	[128, 128, 4]
	BatchNorm1d	[128, 313]	[128, 313]	256	[128]
	ReLU	[128, 313]	[128, 313]		
2. L_1: LSTM: 2		[313, 128] (T)	[313, 128]	132,096	[128, 128]
3. A_1: SoftAttention: 3		[313, 128]	[128]	16,512	[128, 128]
4. B_2: Wave encoder: 1		[1, 128000]	[399, 1024]	variable	
5. L_2: LSTM: 2		[399, 1024]	[399, 128]	590,848	[1024, 128]
6. A_2: SoftAttention: 3		[399, 128]	[128]	16,512	[128, 128]
7. D: Dense layers: 4	Linear	[256]	[256]	65,536	[256, 256]
	ReLU	[256]	[256]		
	Dropout	[256]	[256]		
	Linear	[256]	[6]	1536	[256, 6]
	LogSoftmax	[6]	[6]		
Non-trainable params: wave encoder params					
Trainable params: 1,038,982					
Total params: wave encoder params + 1,038,982					

an accuracy of 68%, being particularly relevant since the FuzzyGCP model is based on another approach for input combination with multiple audio representations. Our best MeWEHV model has an improvement of up to 43.53% over the result achieved by FuzzyGCP, with a VoxForge subset inspired by the one used in the FuzzyGCP paper.

On the YouSpeakers204 and VoxCeleb1 datasets, the best speaker identification model is again MeWEHV using WavLM-large, yielding accuracies of 89.22% and 70.62%, respectively. The largest improvements were achieved with the HuBERT-large-based MeWEHV model with respect to the HuBERT-large model and represent an increase of 88.27% and 58.69%, respectively.

In contrast, for the VBHIR dataset, the best result was obtained with the MeWEHV model using XLSR-Wav2Vec2 as wave encoder, reaching a 94.67% of accuracy. While the highest improvement is 24,23% and was achieved with the MeWEHV-WavLM-large, with respect to the WavLM-large model. It is interesting to note that with this dataset, the models in their base version achieve better results than in their large versions. This may be because in this specific case, being a dataset with relatively few training examples, a smaller number of parameters may reduce the probabilities of overfitting during training.

In the case of the accent identification task, the best model on the Common Voice dataset and NISP dataset is MeWEHV based on WavLM-large, and on the LASC dataset is the MeWEHV model based on XLSR-Wav2Vec2, which achieve accuracies of 42.55%, 84.42%, and 81.59%, respectively. On Common Voice the highest improvement is 20.38% and was obtained with the MeWEHV-HuBERT-base model with respect to HuBERT-base. On NISP dataset the highest improvement is 20.38% and was yielded with the MeWEHVHuBERT-base model with respect to HuBERT-base. Meanwhile, on the LASC dataset the highest improvement is 20.18% and was achieved with MeWEHV-HuBERT-large with respect to HuBERT-large.

The experiments showed a large increase in accuracy for multiple speech classification tasks when MeWEHV is used. The MeWEHV models can be considered a fusion between the wave encoders and the CNNMFCC model, therefore, we can notice that the fusion of both approaches significantly exceeds the performance of each approach separately and that the resulting model takes advantage of the modeling capabilities of both.

TABLE 3. Details of the baseline architectures based on wave encoder branches, adapted to function as classification models, assuming a task with six possible output classes. The number of wave encoder parameters depends on the model chosen for that block, while we set a fixed size of 1024 for the generated embeddings, which is the wave encoder output. The blocks B_2, L_2, and A_2 in the CNNMFCC architecture share the same layers and hyperparameters as the corresponding layers in the *wave encoder branch* of the MeWEHV model.

Layer blocks (index. id: type)	Layers	Input shape	Output shape
1. B_2: Wave encoder		[1, 128000]	[399, 1024]
2. L_2: LSTM		[399, 1024]	[399, 128]
3. A_2: SoftAttention		[399, 128]	[128]
4. D_2: Dense layers	Linear	[128]	[128]
	ReLU	[128]	[128]
	Dropout	[128]	[128]
	Linear	[128]	[6]
	LogSoftmax	[6]	[6]
Non-trainable params: wave encoder params			
Trainable params: 625,542			
Total params: wave encoder params + 625,542			

TABLE 4. The architecture of the baseline CNNMFCC model is described, considering a task with six potential output classes. *(T)* represents that the input of a given layer has been transposed. The blocks B_1, L_1, and A_1 in the CNNMFCC architecture share the same layers and hyperparameters as the corresponding layers in the *MFCC branch* of the MeWEHV model.

Layer blocks (index. id: type)	Layers	Input shape	Output shape
1. B_1: CNN		[128, 641]	[128, 313]
2. L_1: LSTM		[313, 128] (T)	[313, 128]
3. A_1: SoftAttention		[313, 128]	[128]
4. D_1: Dense layers	Linear	[128]	[128]
	ReLU	[128]	[128]
	Dropout	[128]	[128]
	Linear	[128]	[6]
	LogSoftmax	[6]	[6]
Non-trainable params: 0			
Trainable params: 355,654			
Total params: 355,654			

VIII. DISCUSSION

In the proposed MeWEHV architecture, we used two modules that work together to generate rich embeddings. On the one hand, a module for extracting features from raw audios using multiple embedding generation models called *wave encoder*. On the other hand, a module obtains more features using a series of convolutional layers fed by the MFCCs of the original audios, called *MFCC encoder*. The joint work of both modules proved to achieve better results than those obtained by the two modules separately. To validate our approach, we experimented with nine datasets, used for three different tasks, three datasets per task.

Our studies show that optimal results can be obtained after combining both types of inputs in a single architecture and generating rich embeddings. This relates to the findings presented by [51] on FuzzyGCP, where although an approach based on the generation of embeddings was not used, it was shown that combining different audio representations can improve the results obtained with each of these representations individually. In the language identification task with the VoxForge dataset, our approach, based on embeddings generation, proved to be able to achieve better results.

We compared our proposal with five state-of-the-art models and found that the MeWEHV version of each model was able to achieve superior accuracy on all the datasets used. In addition to this, we found that one of the advantages of the proposed model is that only a small number of new parameters are required to be learned to significantly increase the performance of the baseline models.

We can also note that the only model that managed to outperform the MeWEHV-WavLM-large model was the MeWEHV-XLSR-Wav2Vec2 model in the specific cases of accent identification with the LASC dataset and speaker identifications with the VBHIR dataset. This may be because XLSR-Wav2Vec2 was trained to be able to model multiple languages, so it should be able to identify different types of pronunciations and the use of different phonemes, in addition to those used in English, while the other models are only specialized in this language. This can be useful for accent identification, as well as identification of speakers with accents. Also noteworthy is the robustness of the XLSR-Wav2Vec2 model in speaker identification with the YouSpeakers204 dataset, being notably that this, baseline without MeWEHV, has the highest accuracy among all the baselines evaluated, but always lower than our proposal.

In a neural network, the stacking of convolutional layers allows a hierarchical decomposition of the inputs. Because of this, the more convolutional layers are added to a neural network, the higher the level of abstraction that subsequent convolutional layers will achieve.

The MFCC branch of the MeWEHV architecture contains only three convolutional layers. As per the standard structure of a CNN, the first layers are in charge of modeling the low-level features, such as straight lines, edges, and corners, while the later layers are in charge of modeling the high-level features. Based on the results obtained, we can conclude that the addition of low-level features from the MFCC works well complementing the information extracted by the wave encoders.

However, a possible limitation of the MeWEHV architecture is the fact that, since the MFCC branch might not be abstracting high-level features from the inputs due to its relatively low depth., this could imply that the architecture is not exploiting the full potential of the demonstrated complementarity of the evaluated representations.

A future research line would be to evaluate the use of deeper MFCC branches. The exploration could involve replacing the presented convolutional layers with architectures widely used in the field of image processing, such as ResNet, VGG, and DenseNet [68], among others.

TABLE 5. Results in terms of accuracy, obtained in three speech classification tasks, with nine datasets. The models starting with the designation "MeWEHV-X" refer to the models in which the proposed architecture was applied using the baseline model "X" as a wave encoder. The acronym CL refers to the Common Language dataset, LRE17 refers to the NIST 2017 Language Recognition dataset, CV refers to the Common Voice dataset, LASC refers to the Latin American Spanish Corpora dataset, YS204 refers to the YouSpeakers204 dataset, and he acronym VBHIR refers to the *"A Dataset for Voice-Based Human Identity Recognition"* dataset. The best result of each pair of "X" and "MeWEHV-X" models is shown in italics, and the best overall results of each dataset are shown in bold.

Model	**# Params.**	**Language**			**Accent**			**Speaker**		
		VoxForge	CL	LRE17	LASC	CV	NISP	VoxCeleb1	YS204	VBHIR
CNNMFCC	0.36M	67.15%	18.88%	31.88%	84.63%	33.77%	76.16%	41.79%	63.82%	90.67%
XLSR-Wav2Vec2	318.00M	80.70%	37.06%	21.96%	85.39%	33.39%	73.71%	49.18%	83,27%	93.33%
MeWEHV-XLSR-Wav2Vec2	318.42M	*92.70%*	*38.31%*	*32.21%*	**87,62%**	*33.69%*	*83.11%*	*64.40%*	*87,96%*	**94.67%**
HuBERT-base	95.30M	87.09%	39.89%	32.78%	77.43%	32.09%	78.65%	51.66%	59.03%	86.66%
MeWEHV-HuBERT-base	95.72M	*94.19%*	*44.87%*	*33.06%*	*82.68%*	*38.63%*	*82.65%*	*64.86%*	*87.35%*	*91.55%*
HuBERT-large	317.23M	85.90%	46.55%	28.01%	67.89%	34.84%	53.50%	40.87%	46.65%	83.11%
MeWEHV-HuBERT-large	317.65M	*94.49%*	*49.45%*	*33.51%*	*81.59%*	*35.73%*	*82.10%*	*64.86%*	*87.83%*	*87.11%*
WavLM-base	95.32M	92.50%	57.33%	34.55%	83.29%	35.35%	78.70%	47.55%	61.27%	83.55%
MeWEHV-WavLM-base	95.74M	*97.30%*	*62.51%*	*36.40%*	*83.49%*	*36.40%*	*83.83%*	*67.09%*	*87.49%*	*90.22%*
WavLM-large	317.24M	96.89%	63.36%	36.49%	80.20%	38.60%	78.61%	61.49%	67.53%	71.55%
MeWEHV-WavLM-large	317.66M	**97.60%**	**72.53%**	**41.05%**	*83.63%*	**42.55%**	**84.42%**	**70.62%**	**89.22%**	*88.89%*

Another limitation of the present work is that we only evaluated the use of MFCCs as an imposed representation. Although it was demonstrated that MFCCs are functional for the purpose of our work to generate rich embeddings, it also opens the door to a future research line in which the performance of the MeWEHV architecture is compared to new architectures based on other representations, like spectrograms, Power-Normalized Cepstral Coefficients (PNCC) [69], and others.

Related to the previous point, a limitation in this work is that only two acoustic representations were used at the same time. In a future research line, the creation of new architectures based on MeWEHV fed by more than two representations can be explored in order to establish if the addition of further representations increases, even more, the results achieved.

Finally, the presented work demonstrates experimentally that there is a complementarity between the used embeddings generated from raw audio waves and the embeddings generated from MFCCs. A future research line would also consist of determining the causes of this complementarity, establishing what information is *missing* in both representations.

IX. CONCLUSION

In this work we have proposed MeWEHV, a machine learning model architecture that enriches the embeddings, generated by a pre-trained wave encoder, using features extracted from MFCC representations. MeWEHV was tested on the language identification task with the VoxForge, Common Language, and LRE17 datasets, achieving accuracies of up to 97.60%, 72.53%, and 41.05%, respectively, superior to other state-of-the-art approaches. It should be noted that the MeWEHV architecture only requires 1.04M additional parameters in addition to the wave encoder parameters, representing only 0.33% to 1.09% additional parameters.

Furthermore, the model was tested in the identification of accents with the Latin American Spanish Corpora, achieving an accuracy of up to 87.62%. This is the first result reported with this dataset in this specific task, which will allow future research to have a reference result to compare with. Moreover, it was tested on the Common Voice, and NISP datasets achieving an accuracy of up to 42.55%, and 84.42%, respectively. In all three datasets the MeWEHV models achieved the highest results.

We proposed YouSpeakers204, a new speaker identification dataset, highly balanced by accent and speaker gender, in which MeWEHV obtained 89.22%, which is the highest accuracy. Together with the dataset, we proposed training, test, and validation sets, which can be used by other researchers for a fair comparison. In speaker identification, we also tested the VoxCeleb1 dataset, and the VBHIR dataset, obtaining the best results with MeWEHV, with accuracies of up to 70.62%, and 94.67%, respectively.

In all experiments, the results of MeWEHV models were compared with the CNNMFCC model which is an MFCC-fed CNN model, and with five state-of-the-art embedding generation models, i.e. WavLM base, WavLM large, XLSR-Wav2Vec2, HuBERT base and HuBERT large, outperforming, in all cases, their results. In this way, we demonstrated that our approach is superior to all baselines, in multiple speech classification tasks.

Thus, this work allows the use of a machine learning architecture that requires training with a relatively low additional computational cost and consistently achieves superior results than the baselines. Our architecture provides a general framework that can be used with other pre-trained models as wave encoders.

This paper demonstrates experimentally that there is a complementarity between the information that MeWEHV is able to extract from the embeddings generated by a pre-trained model from raw audio and the MFCCs extracted from the same audios, since the results of using both representations outperformed the results of using each representation separately.

ACKNOWLEDGMENT

This publication reflects the views only of the authors, and the European Union's Horizon 2020 Research and Innovation Framework Programme, H2020 SU-FCT-2019 cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

[1] A. B. Nassif, I. Shahin, S. Hamsa, N. Nemmour, and K. Hirose, "CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions," *Appl. Soft Comput.*, vol. 103, May 2021, Art. no. 107141.

[2] H. Huang, X. Xiang, Y. Yang, R. Ma, and Y. Qian, "AISpeech-SJTU accent identification system for the accented English speech recognition challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 6254–6258.

[3] Y. Nie, J. Zhao, W. Zhang, and J. Bai, "BERT-LID: Leveraging BERT to improve spoken language identification," in *Proc. 13th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, K. A. Lee, H. Lee, Y. Lu, and M. Dong, Eds., 2022, pp. 384–388.

[4] J. S. Garofolo et al., "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," Web Download, Linguistic Data Consortium, Philadelphia, PA, USA, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93s1, doi: 10.35111/17gk-bn40.

[5] K. MacLean. (2018). *Voxforge*. [Online]. Available: http://www.voxforge.org/

[6] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. 12th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5206–5210.

[8] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for ASR with limited or no supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7669–7673.

[9] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T. N. Sainath, and A. Sethy, "Developing speech recognition systems for corpus indexing under the IARPA Babel program," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 6753–6757.

[10] A. Guevara-Rukoz, I. Demirsahin, F. He, S. C. Chu, S. Sarin, K. Pipatsrisawat, A. Gutkin, A. Butryna, and O. Kjartansson, "Crowdsourcing Latin American Spanish for low-resource text-to-speech," in *Proc. 12th Lang. Resour. Eval. Conf. (LREC)*, 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, and S. Piperidis, Eds. Paris, France: European Language Resources Association, 2020, pp. 6504–6513.

[11] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Stockholm, Sweden, F. Lacerda, Ed., Aug. 2017, pp. 2616–2620.

[12] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. 21st Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, H. Meng, B. Xu, and T. F. Zheng, Eds., 2020, pp. 2757–2761.

[13] Q. Wang, Y. Yu, J. Pelecanos, Y. Huang, and I. Lopez-Moreno, "Attentive temporal pooling for conformer-based streaming language identification in long-form speech," in *Proc. Speaker Lang. Recognition Workshop (Odyssey)*, 2022, pp. 255–262.

[14] A. B. Nassif, I. Shahin, A. Elnagar, D. Velayudhan, A. Alhudhaif, and K. Polat, "Emotional speaker identification using a novel capsule nets model," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116469.

[15] D. Wang, S. Ye, X. Hu, S. Li, and X. Xu, "An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model," in *Proc. 22nd Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brno, Czechia, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds., Aug./Sep. 2021, pp. 3266–3270.

[16] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," in *Proc. 21st Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, H. Meng, B. Xu, and T. F. Zheng, Eds., 2020, pp. 140–144.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020, pp. 12449–12460.

[18] W.-N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.

[19] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[20] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. INTERSPEECH*, Aug. 2021, pp. 2426–2430.

[22] D. Berrebbi, J. Shi, B. Yan, O. López-Francisco, J. Amith, and S. Watanabe, "Combining spectral and self-supervised features for low resource speech recognition and translation," in *Proc. INTERSPEECH*, Sep. 2022, pp. 3533–3537.

[23] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Dec. 2014, pp. 3581–3589.

[24] J. Camacho-Collados and M. T. Pilehvar, "From word to sense embeddings: A survey on vector representations of meaning," *J. Artif. Intell. Res.*, vol. 63, pp. 743–788, Dec. 2018.

[25] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, A. Moschitti, B. Pang, and W. Daelemans, Eds., Oct. 2014, pp. 36–45.

[26] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognit. Artif. Intell.*, vol. 116, no. 1, pp. 374–388, 1976.

[27] I. Kamarulafizam, S.-H. Salleh, J. Najeb, A. Ariff, and A. Chowdhury, "Heart sound analysis using MFCC and time frequency distribution," in *Proc. World Congr. Med. Phys. Biomed. Eng.* Berlin, Germany: Springer, 2007, pp. 946–949.

[28] Z. Li, M. Zhao, Q. Hong, L. Li, Z. Tang, D. Wang, L. Song, and C. Yang, "AP20-OLR challenge: Three tasks and their baselines," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Auckland, New Zealand, Dec. 2020, pp. 550–555.

[29] D. Wang and X. Zhang, "THCHS-30: A free Chinese speech corpus," 2015, *arXiv:1512.01882*.

[30] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. 5th Eur. Conf. Speech Commun. Technol. (EUROSPEECH)*, Rhodes, Greece, G. Kokkinakis, N. Fakotakis, and E. Dermatas, Eds., Sep. 1997, pp. 1743–1746.

[31] I. Shahin, A. B. Nassif, and S. Hamsa, "Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments," *Neural Comput. Appl.*, vol. 32, no. 7, pp. 2575–2587, Apr. 2020.

[32] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[33] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Graz, Austria, G. Kubin and Z. Kacic, Eds., Sep. 2019, pp. 814–818.

[34] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, "The accented English speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 6918–6922.

[35] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[36] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 776–780.

[37] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.

[38] S. Yang, P. Chi, Y. Chuang, C. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G. Lin, T. Huang, W. Tseng, K. Lee, D. Liu, Z. Huang, S. Dong, S. Li, S. Watanabe, A. Mohamed, and H. Lee, "SUPERB: Speech processing universal performance benchmark," in *Proc. 22nd Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brno, Czechia, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds., Aug./Sep. 2021, pp. 1194–1198.

[39] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020.

[40] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd Annu. ACM Conf. Multimedia Conf. (MM)*, Brisbane, QLD, Australia, X. Zhou, A. F. Smeaton, Q. Tian, D. C. A. Bulterman, H. T. Shen, K. Mayer-Patel, and S. Yan, Eds., Oct. 2015, pp. 1015–1018.

[41] C. Kroos, O. Bones, Y. Cao, L. Harris, P. J. B. Jackson, W. J. Davies, W. Wang, T. J. Cox, and M. D. Plumbley, "Generalisation in environmental sound classification: The 'making sense of sounds' data set and challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 8082–8086.

[42] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. Scenes Events Workshop (DCASE)*, 2018, pp. 9–13.

[43] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Proc. Scenes Events Workshop (DCASE)*, Nov. 2018, pp. 69–73.

[44] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[45] M. Mulimani and S. G. Koolagudi, "Acoustic event classification using spectrogram features," in *Proc. IEEE Region Conf. (TENCON)*, Jeju, South Korea, Oct. 2018, pp. 1460–1464.

[46] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019.

[47] Sarthak, S. Shukla, and G. Mittal, "Spoken language identification using ConvNets," in *Proc. 15th Eur. Conf. Ambient Intell.*, in Lecture Notes in Computer Science, Rome, Italy, vol. 11912, I. Chatzigiannakis, B. E. R. de Ruyter, and I. Mavrommati, Eds. Cham, Switzerland: Springer, Nov. 2019, pp. 252–265.

[48] R. A. Lee and J. R. Jang, "A syllable structure approach to spoken language recognition," in *Proc. 6th Int. Conf. Stat. Lang. Speech Process. (SLSP)*, in Lecture Notes in Computer Science, Mons, Belgium, T. Dutoit, C. Martín-Vide, and G. Pironkov, Eds., vol. 11171. Cham, Switzerland: Springer, Oct. 2018, pp. 56–66.

[49] K. S. Ahmad, A. S. Thosar, J. H. Nirmal, and V. S. Pande, "A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network," in *Proc. 8th Int. Conf. Adv. Pattern Recognit. (ICAPR)*, Kolkata, India, Jan. 2015, pp. 1–6.

[50] A. Meghanani, A. C. S., and A. G. Ramakrishnan, "An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Shenzhen, China, Jan. 2021, pp. 670–677.

[51] A. Garain, P. K. Singh, and R. Sarkar, "FuzzyGCP: A deep learning architecture for automatic spoken language identification from speech signals," *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114416.

[52] K. Prahallad, N. K. Elluru, V. Keri, R. S, and A. W. Black, "The IIIT-H indic speech databases," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Portland, OR, USA, Sep. 2012, pp. 2546–2549.

[53] A. Baby, A. L. Thomas, N. Nishanthi, and T. Consortium, "Resources for Indian languages," in *Proc. Text, Speech Dialogue*, 2016, pp. 1–8.

[54] M. Z. Boito, W. Havard, M. Garnerin, É. L. Ferrand, and L. Besacier, "Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible," in *Proc. 12th Lang. Resour. Eval. Conf. (LREC)*, Marseille, France, 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, and S. Piperidis, Eds. Paris, France: European Language Resources Association, May 2020, pp. 6486–6493.

[55] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Adelaide, SA, Australia, Apr. 1994, pp. 305–308.

[56] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 92–97.

[57] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5329–5333.

[58] L. Gao, K. Xu, H. Wang, and Y. Peng, "Multi-representation knowledge distillation for audio classification," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 5089–5112, Feb. 2022.

[59] B. Zhu, K. Xu, Q. Kong, H. Wang, and Y. Peng, "Audio tagging by cross filtering noisy labels," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2073–2083, 2020.

[60] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, Amsterdam, The Netherlands, vol. 9911, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, Oct. 2016, pp. 499–515.

[61] H. Yao, D.-l. Zhu, B. Jiang, and P. Yu, "Negative log likelihood ratio loss for deep neural network classification," in *Proc. Future Technol. Conf. (FTC)*, vol. 1. Cham, Switzerland: Springer, 2020, pp. 276–282.

[62] G. Sinisetty, P. Ruban, O. Dymov, and M. Ravanelli, "CommonLanguage," Zenodo, Version 0.1, Jun. 2021, doi: 10.5281/zenodo.5036977.

[63] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. 12th Lang. Resour. Eval. Conf. (LREC)*, Marseille, France, 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, and S. Piperidis, Eds. Paris, France: European Language Resources Association, May 2020, pp. 4218–4222.

[64] S. O. Sadjadi, T. Kheyrkhah, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "Performance analysis of the 2017 NIST language recognition evaluation," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Hyderabad, India, B. Yegnanarayana, Ed., Sep. 2018, pp. 1798–1802.

[65] S. B. Kalluri, D. Vijayasenan, S. Ganapathy, R. R. M, and P. Krishnan, "NISP: A multi-lingual multi-accent dataset for speaker profiling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 6953–6957.

[66] A. Ahamad, A. Anand, and P. Bhargava, "AccentDB: A database of non-native English accents to assist neural speech recognition," in *Proc. 12th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, May 2020, pp. 5353–5360.

[67] A. Baha'A, H. S. A. Arja, B. Y. Maayah, and M. M. Al-Taweel, "A dataset for voice-based human identity recognition," *Data Brief*, vol. 42, Jan. 2022, Art. no. 108070.

[68] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020.

[69] C. Kim and R. M. Stern, ''Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction,'' in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brighton, U.K., Sep. 2009, pp. 28–31.

ANDRÉS CAROFILIS received the B.S. degree in computer engineering from the Central University of Ecuador, Ecuador, in 2018, and the M.Sc. degree in logic, computation, and artificial intelligence from the University of Seville, Spain, in 2019. He is currently pursuing the Ph.D. degree with Universidad de León, Spain. He is currently a Researcher with Universidad de León. His current research interests include speech processing, pattern recognition, data science, and computer vision.

LAURA FERNÁNDEZ-ROBLES received the M.Sc. degrees in intelligent systems in engineering and industrial engineering from Universidad de León, Spain, in 2009, and the joint Ph.D. degree from the University of Groningen, The Netherlands and Universidad de León, in 2016. She has been an Associate Professor and a Researcher with Universidad de León, since 2021, where she started as a Lecturer, in 2012. Her current research interests include computer vision, pattern recognition, and data science applied to industrial, cybersecurity, medical, and animal ethology problems.

ENRIQUE ALEGRE received the M.Sc. degree in electrical engineering from the University of Cantabria, in 1994, and the Ph.D. degree from Universidad de León, Spain, in 2000. He is currently the Head of the Research Group for Vision and Intelligent Systems (GVIS) and a Full Professor with the Department of Electrical, Systems and Automation Engineering, Universidad de León. His research interests include computer vision and machine learning in general and also deep learning and natural language processing, specially oriented to cybersecurity, crime control, and prevention problems.

EDUARDO FIDALGO received the M.Sc. and Ph.D. degrees in industrial engineering from Universidad de León, in 2008 and 2015, respectively. He is currently an Assistant Professor with the Group for Vision and Intelligent Systems (GVIS), whose main objective is researching and developing solutions to cybersecurity and cybercrime-related problems using artificial intelligence. His current research interests include natural language processing, computer vision, and machine and deep learning.