ISSN 2007-9737

Spoken Language Identification for Short Utterance with Transfer Learning

Ana Montalvo-Bereau*, Jose Ramón Calvo-de-Lara,
Gabriel Hernández-Sierra, Flavio Reyes-Díaz

Centro de Aplicaciones de Tecnologías de Avanzada,
Cuba

{amontalvo, jcalvo, gsierra, freyes}@cenatav.co.cu

Abstract. Spoken language recognition is a research field that has received considerable attention due to its impact on several tasks related to multilingual speech processing. While it has been demonstrated that the use of contextual and auxiliary task information can enhance the results within this field, this avenue has not been fully explored. In the present work, we propose to address the spoken language recognition task in short utterances by considering two speech-related tasks as auxiliaries in a multi-tasking architecture. The primary task was language recognition, with sex and speaker identity serving as auxiliary tasks. Three models from disparate approaches were implemented and trained in a single-task and multi-task learning paradigm. The models considered were 2D-CNN based, one of which was a proposed configuration designed to address less than a second utterances. The experiments were conducted on a subset of the VoxForge corpus, with a markedly limited amount of signals. The results demonstrate that the spoken language recognition task benefits from multi-task learning by using sex and speaker identity as auxiliary tasks over three different models.

Keywords. Spoken language recognition, deep learning, transfer learning, multi-task learning.

1 Introduction

Spoken language recognition (SLR) is the automated process of identifying the language spoken in a speech sample. It serves as a foundational technology for various multilingual speech-processing applications.

One challenge for voice assistant products is the requirement for predetermined language usage by the user, either explicitly declared or automatically recognized by the system. A reduction in the time required to identify the spoken language in a signal would enhance the user experience, particularly in tasks such as spoken term detection, speech-to-text transcription, and automatic spoken language translation.

Minimizing latency and processing time in spoken language recognition, especially in online mode, is crucial, making the length of the input segment for inference and the size of the model key parameters. At present, SLR systems rely on deep learning models, either in the stage of feature extraction by learning representations [18], or in end-to-end architectures that jointly model feature extraction and classification of the system [2].

Currently, deep neuronal networks with end-to-end architecture lead SLR, especially for short utterances (3 seconds) [23]. Compared with long utterances, the feature representation of short utterances has a large variation, which prevents the model from generalizing well. The challenge of improving the generalization of the model on short utterances remains.

Most machine learning techniques are narrowly focused and trained in isolation with a single task . This approach, which we will refer to as single-task learning, neglects certain fundamental aspects of human learning.

ISSN 2007-9737

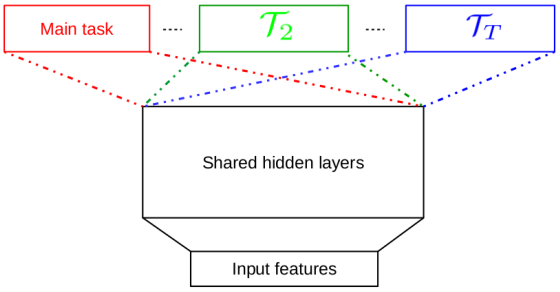

Fig. 1. Hard parameter sharing MTL architecture with one main task and $T-1$ auxiliary tasks

Humans enter each new learning task with the knowledge acquired from prior learning experiences. Yet, human learning frequently entails addressing multiple learning tasks concurrently. The set of techniques within machine learning that allows joint learning of related problems is called multi-task learning (MTL [3]).

The basic idea of MTL is to improve the learning of a main task through the use of the information contained in the training signals of other tasks called auxiliary and related to the main one, using a shared representation. This is based on the assumption that what is learned for each task can help other tasks to be learned better. MTL improves generalization by drawing on information contained in related tasks [3].

Speech, as a component of the complex voice signal, conveys semantic information on the message it transmits, but also several non-semantic characteristics, including the speaker's identity, sex, age, language, accent, emotional and health state, and so forth [24].The understanding of these characteristics by automatic learning systems has two main advantages.

First, it could help prevent bias and discrimination in voice applications based on artificial intelligence. Secondly, some studies support the notion that the joint modeling of this information has the potential to positively influence spoken language classification [8]. State-of-the-art approaches, entirely based on deep neural networks, have demonstrated impressive performance for short utterances in SLR [22]. Despite the progress that has been made, these techniques are susceptible to overfitting the training set or domain generalization problems [1].

In this paper, the use of MTL in SLR is evaluated, considering non-semantic tasks, using three different approaches. A preliminary study on this subject was conducted previously [16], and the present work constitutes an extension of that study. The former was the first attempt to assess SLR based on non-semantic multi-task learning, starting from the knowledge transferred from the real images domain to audio classification tasks.

In this study, two additional approaches are considered to test the hypothesis that the inclusion of auxiliary information, such as the identity of the speaker and its sex, could enhance SLR performance for short and very short utterances (less than 3 seconds).

The objective is to experimentally verify that the proposed auxiliary tasks (speaker's identity and sex) for SLR yield benefits of multi-task learning, particularly for very short-duration signals. This is the main contribution of the present work.

The remainder of this paper is organized as follows: section 2 is devoted to a review of the methods required to build a 2D-CNN multi-tasking setup and its applications to SLR. The experimental protocol and dataset used during experimentation are described in section 3, as well as a description of the models assessed. The results are presented and discussed in section 4. Finally, the final conclusions of this work are summarized in 5.

2 Methods

The performance metric used is accuracy, defined as the ratio of correctly classified samples (prediction) to the total number of predictions. The value of this metric ranges between 0 and 1 and is expressed as a percentage (%):

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}}. \quad (1)$$

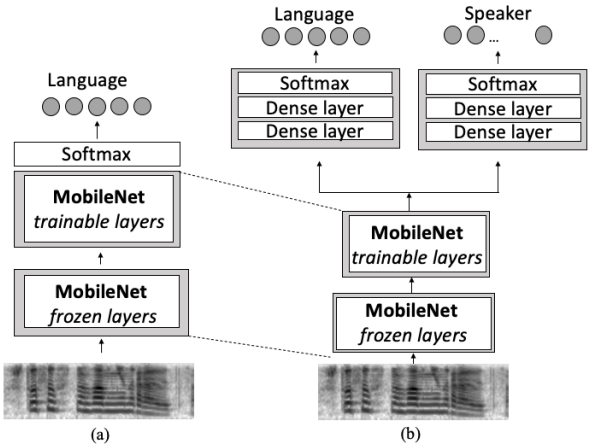

Fig. 2. Scheme of MobileNet used in single-task (a) and multi-task (b) setups

2.1 Multi-task Learning

Multi-task learning is a collection of techniques intended to learn multiple tasks simultaneously instead of learning them separately [3]. Beyond the biological and pedagogical motivations, from the machine learning perspective, learning multiple related tasks leads to inductive bias, which helps the models generalize better.

This generalization capacity represents a significant aspect when little training data is available and in facing very short-duration test utterances, as addressed in the last model proposed by the authors in the present paper. The concept of multi-task learning (MTL) has been a topic of interest for some time. The rationale behind this approach is that if two related tasks are present, then the learned features should be related as well.

This is particularly relevant in the context of deep models, such as deep convolutional neural networks (CNNs), due to the presence of a hierarchy of features [17].Even if the highest-level features are task-specific, lower-level features can likely be shared. This approach offers the advantage of augmenting the data set for those lower-level features by training them jointly on related tasks.

The two most commonly used methods for performing MTL in deep neural networks are either hard or soft parameter sharing of hidden layers.Hard parameter sharing is the most commonly used [3] and involves sharing the hidden layers between all tasks while maintaining several task-specific output layers. The sharing of representations between related tasks enables the model to generalize better on the main task, thereby reducing the risk of overfitting.

In soft parameter sharing, each task has its own model, and the learning process penalizes the distance between the different parameters. Unlike hard sharing, this approach provides greater flexibility for the tasks by only loosely coupling the shared space representations.

In situations where it is necessary to solve several classification tasks simultaneously, MTL is an optimal solution. However, in most situations, the focus is on the performance of a single task.

The relationship between tasks is a crucial aspect of MTL, although it is challenging to assess [29]. The application of MTL for SLR has primarily focused on relating the phonetic information with the language, either in end-to-end approaches [10] as during the representation stage [29].

Additionally, there are instances where the recognition of language or dialect serves as an auxiliary task, with the primary objective being to relate it to phonemes to enhance the efficacy of speech recognition [13].

In the case of negatively correlated tasks, such as language and domain differences, adversarial MTL has been employed to develop a model that is less reliant on the domain [1]. Some theoretical advances have been made in understanding task-relatedness ; however, there has been limited progress towards this goal.

To compensate for this, researchers have recently explored MTL from a more experimental point of view, correlating performance gains with task properties to achieve a better understanding of when models can profit from auxiliary tasks [10, 15].

This paper sheds light on the specific task relations that can lead to gains from MTL models over single-task setups in SLR. To the best of the authors' knowledge, the consideration of speakers'

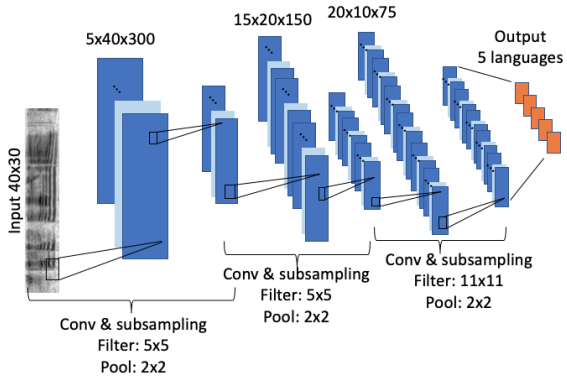

Fig. 3. Representation of Lozano's single-task model architecture, with 3 hidden layers of 5, 15 and 20 filters respectively to discriminate among 5 languages

identity and sex as auxiliary tasks for spoken language recognition in a multi-tasking setup has not been explored before.

2.1.1 Auxiliary Tasks

It is not reasonable to assume that information gathered through the learning of a set of tasks will be relevant to the learning of another task that has nothing in common with the already learned set of tasks. From an engineering perspective, speech recognition and speaker recognition are independent tasks.

However, the human brain interprets and decodes information from both speaker traits and linguistic content from speech in a joint corroborative manner [8]. Similarly, unified frameworks for speaker and language recognition have been attempted using a shared deep neural network, which outperforms the single-task implementation.

On the one hand, language and speaker recognition tasks share numerous common techniques, including cepstral feature extraction and well-established Gaussian-based modeling. On the other hand, researchers in both areas have a history of learning from each other. For instance, the success of i-vector [5] and x-vector [26] representations originally proposed for speaker recognition has been immediately transposed to language recognition [25]. Several technologies are shared between speakers and language recognition. Consequently, the proposed ideas in one application can be also used in another. In [4] speaker identity and sex have been demonstrated to be correlated, thereby establishing a link between these non-semantic tasks and the identity of the speaker.

As for selecting which auxiliary tasks to employ, these studies have strongly encouraged us to explore the benefits of using sex and speaker identity to the SLR main task. The rationale for employing both auxiliary tasks is to direct the networks' attention to the correlation between the variability of language posteriors and two of the speaker attributes. If the system can differentiate the speaker's characteristics, then this information can be utilized for a more accurate interpretation of the distortion introduced by one speaker in comparison to another.

2.1.2 Multi-task Architecture

The multi-task architecture employs hard parameter sharing, a technique originally proposed by [3] that has remained the norm for almost 30 years. Figure 1 represents a hard parameter sharing MTL model, with one main task and $T-1$ auxiliary tasks, where T is the number of tasks. There are two fundamental aspects shared between these MTL systems.

First, all tasks use the same base representation of the input data. Second, the task-specific portions of the network all begin with the same representation from the final shared layer, as all the tasks share the same network weights until bifurcation. In this setup, each task contributes to the cost function with its own individual loss, as illustrated in equation 2.

For the sake of simplicity, lets consider T tasks, denoted as $\mathcal{T}_1, \mathcal{T}_2 ... \mathcal{T}_T$. The training data of each task is represented as $\mathcal{D}^{\mathcal{T}_t}$ where $t \in \{1, 2, ..., T\}$. Being $\mathcal{L}_{\mathcal{T}_t}(\mathcal{D}^{\mathcal{T}_t}, \theta)$ the loss function of the task $\mathcal{T}_t$ and θ the total parameters of the MTL model, the objective is to estimate the model parameters θ^* such that:

$$\theta^* = \arg\min_{\theta} \sum_{t=1}^{T} \lambda^{\mathcal{T}_t} \mathcal{L}_t(\mathcal{D}^{\mathcal{T}_t}, \theta), \qquad (2)$$

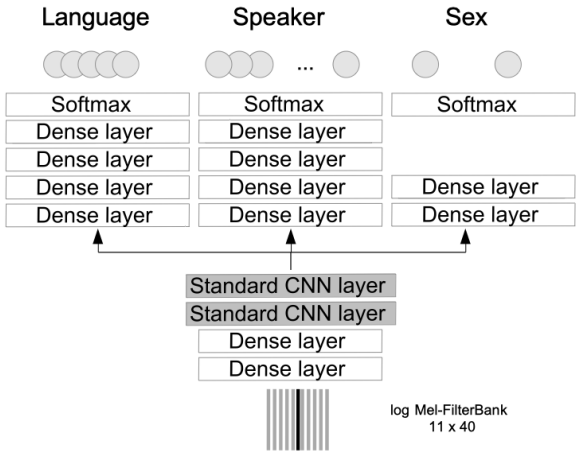

Fig. 4. Representation of multi-task frame level model

where $\lambda^{\mathcal{T}_t}$ is the non-negative weight, which considers the impact the particular task $\mathcal{T}_t$ will have on the estimation of the system parameters.

In most cases, the auxiliary tasks are dropped at test time, keeping only the main task outputs. For single-task setup, only the main task remains.

2.2 2D-CNN

The three models assessed are CNN-based. Deep convolutional neural networks (CNNs) have been demonstrated to be effective in reducing spectral variations and modeling correlations in acoustic features [28], and have been utilized in a multitude of audio classification tasks [9]. There exist CNN-based approaches that utilize raw or minimally preprocessed audio as an input, employing one-dimensional convolutions. However, the majority of outcomes have been achieved through the utilization of two-dimensional (2D) CNNs on spectrograms [19].

In general, deep CNN models comprise multiple convolution layers connected to one or several fully connected layers. The convolution layers may be regarded as feature extraction layers, while the fully connected layers may be considered as classification layers. To feed the 2D-CNN-based models, acoustic features were extracted from the audio signals and presented in the time domain as spectrograms.

The acoustic features were computed using Kaldi [20]. The signal is subjected to a pre-emphasis filter, then divided into 20 millisecond frames (with 10 ms overlap), and a window function is applied to each frame. Subsequently, a short-time Fourier transform is computed to obtain the power spectrum.

Subsequently, a triangular Mel scale filter bank and the logarithm of the energy output of the individual bandpass filter are applied, resulting in a representation of 40-dimensional vectors known as a spectrogram or log Mel filter bank (log Mel-FBank). MFCCs [6] are the most widely used input features for speech analysis tasks.

To obtain MFCCs, a discrete cosine transform is applied to the log Mel-FBank, retaining a number of the resulting coefficients while discarding the rest. It has been found that the last step removes information and destroys spatial relations; therefore, it is usually omitted, which yields the log Mel-FBank output, a popular feature across the speech community.

3 Experimental Setup

3.1 Corpus

The corpus utilized was VoxForge [12], a free and open-source corpus of voices containing samples of more than 18 different languages. The data consisted of audio files of approximately 5 to 10 seconds in duration, accompanied by the transcription of the spoken text, as well as labels related to the language, sex, and identity of the speaker.

As the corpus is comprised of audio samples submitted by individuals from diverse geographical and linguistic backgrounds, the quality of the samples varies according to the recording conditions and equipment used by the contributors.

This results in a significant variation of speech quality between samples, which is representative of real-world scenarios. For experimentation, we defined a VoxForge subset consisting of five languages: German, Spanish, French, English, and Russian. Approximately 38 minutes per language were allocated for the creation of training, validation, and test sets, ensuring a balance of

Table 1. Language accuracy comparison of single and multi-task architectures using MobileNet base model

Setup	Acc train (%)	Acc val (%)	Acc test (%)
STL	100	80.9	76.5
MTL	99.7	82.86	83.42

Table 2. Language accuracy comparison of single and multi-task architectures using Lozano's model adaptation

Setup	Acc train (%)	Acc val (%)	Acc test (%)
STL	96.67	93.92	91.39
MTL	98.24	97.65	96.44

female and male speakers. Although this is not an objective pursued in this research, the defined scenario has a very low data-resources profile compared to other speech-related applications.

3.2 Models Description

The influence of MTL on spoken language recognition across three distinct models is evaluated. The main task for all models is language recognition. The first model introduces speaker identity as auxiliary task, while the next two models employ both speaker and sex identification tasks as auxiliaries.

Whereas the initial two models (MobileNet pretrained model and Lozano's model) utilize short utterances (3-second duration samples), the proposed third model employs very short duration utterances (110 ms). The parameter tuning for each of the investigated methods was conducted over the validation set.

3.2.1 MobileNet Pre-trained Model

The first SLR approach evaluated is an end-to-end architecture based on a CNN pre-trained on a set of images [21], for a more detailed explanation, please refer to previous work [16]. This technique of transferring knowledge is known as transfer learning. In this case, the parameters of the initial network are trained on images of real objects that are very different from the spectrograms that constitute the input to this method.

As input, we considered the use of spectrograms of the audio signal. This two-dimensional representation in the time-frequency domain can be considered as an image. The 300 initial temporal vectors from the spectrogram of each signal were concatenated to form a 40×300 feature matrix, which was subsequently transformed into a gray-scale image. Only the first three seconds of audio from each signal were utilized after the initial silence segments were eliminated.

To initialize the network, the MobileNetV2 [21] was employed as a pre-trained model. MobileNetV2 was developed by Google and trained on the Imagenet data set, which comprises 1.4 million images and 1,000 types of web images. The proposed model comprises the MobileNetV2 with 53 layers of depth, of which the initial 30 remain fixed or "frozen" and the remaining network layers are re-trained (see Figure 2). The initial block will be referred to as pre-trained and can be considered a high-level abstraction features extractor, which transforms the input sequence into a characteristics map.

The pre-trained block is succeeded by trainable layers of the convolutional block, which are also part of the MobileNetV2. However, unlike the pre-trained block, their parameters are learned over the training set. In the multi-task setup, with MobileNet's trainable layers, two individual task branches are trained: language as the main task and speaker identity as the auxiliary task [16].

The layers of the individual branches are dense layers with 512 neurons, using the ReLU activation function and a 25% dropout. The Adam optimizer algorithm was employed, as well as a learning rate of 10^{-4}. It is important to note that the estimation of the model parameters is directly affected by the knowledge embedded in the frozen layers, which is transferred to the trainable layers that follow it.

3.2.2 Lozano's Model

The model proposed by [11] was one of the first attempts to build an SLR system intended to deal with short utterances using CNNs as end-to-end approach. In the referenced paper, the network was fed with speech segments of three seconds, in

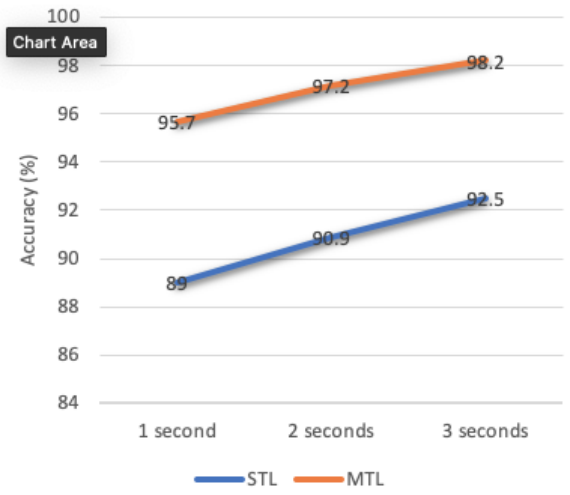

Fig. 5. Accuracy with posteriors combination of frame-level model using majority voting for language recognition task over test set

the form of a matrix of dimensions 56×300, formed by 300 frames.Each frame is represented with a vector of 56 Mel Frequency Cepstral Coefficients and its derivatives (MFCC-SDC) [27]. The CNN system proposed then, obtained comparable results to the i-vector approach, having much less free parameters.

The CNN system proposed then obtained comparable results to the i-vector approach, having much less free parameters. We have replicated Lozano's network architecture, modifying the dimensions of the input and output layers (see Figure 3 for the modified version). This results in a more streamlined configuration than the MobileNet pre-trained model in terms of the number of parameters to estimate, although it still faces the challenge of recognizing speech signals of at least three seconds in duration.

In the adaptation of Lozano's model presented in this work, instead of MFCC-SDC acoustic features, log Mel-FBank features composing the spectrogram were utilized. It has been demonstrated that convolution on mel spectrograms is more beneficial than on decorrelated coefficients [14].

Each speech signal was segmented into intervals of 3 seconds, with an overlap of 50%. The aforementioned segments were represented by a matrix of dimensions 40×300 formed with the log Mel-FBank features and were used to feed the 2D-CNN models.

Details about STL Lozano's model can be seen in Figure 3, where 3 hidden layers of 5, 15, and 20 filters define the architecture as proposed in the original paper. In the case of the multi-task setup, the output of the third convolutional layer was flattened and passed on to the three branches in parallel.

The language, speaker, and sex branches are formed with two dense layers of 256 and 128 neurons, respectively, a dropout factor of 25%, and a softmax activation function at the end. Regarding the training of the network, the algorithm used is stochastic gradient descent with a learning rate of 0.01 and based on minibatches of 32 samples.

3.2.3 Frame-level Model

The SLR frame-level approach, which was initially proposed in [7], demonstrates how deep neural networks are particularly well-suited for SLR in real-time applications. This is due to their capacity to emit a language identification posterior at each new frame of the test utterance. In [7], the authors compare their proposal with the i-vector approach using very short test utterances ($\leq 3s$).

In contrast, our study aims to investigate the potential contribution of MTL to the system under similar test duration conditions. Our proposal focuses on SLR with very short utterances, less than 3 seconds speech segments. Each frame is represented with its log Mel-FBank feature vector, and spliced together with 5 left and 5 right context features to form a 40×11 dimensional feature matrix.

The spliced features are fed as input to the 2D-CNN, thus enabling the model to be trained with samples corresponding to a temporal context of $11 \times 0.01 = 0.11$ seconds. The spliced features are fed as input to the 2D-CNN, so the model is trained with samples corresponding to a context of $11 \times 0.01 = 0.11$ seconds.

Table 3. Language accuracy comparison of single and multi-task architectures using frame-level model

Setup	Acc train (%)	Acc val (%)	Acc test (%)
STL	95.39	79.08	78.35
MTL	97	88.82	88.11

Table 4. SLR multi-task models' relative improvement over test set. Accuracy(%) reported

Model	STL	MTL	Improvement
Mobile	76.50	83.42	6.92
Lozano	91.39	96.44	5.05
Frame level	92.50	98.20	5.70

As illustrated in Figure 4, the lower-level representations of the network across the tasks are shared. The first pair of layers is fully connected with 256 neurons, followed by a convolutional layer comprising a convolution and max-pooling operation. The convolution operation is achieved by weight sharing across the entire training sample. A total of 256 filters and ReLU activations are employed, with each 2D convolution operated with zero-padding using a 3×3 kernel size.

A second convolutional layer is utilized to model the local spectro-temporal correlations of the speech spectrogram, this time using 128 filters of the same size. The resulting feature map is then flattened and passed on to dense layers for language, speaker, and sex mapping, as illustrated in Figure 4 of the MTL setup.

In the MTL setup, the individual branches comprise four dense layers for language and speaker classification and two dense layers for the binary task of sex classification. The STL setup employs a configuration similar to Figure 4 which is designed to ignore the speaker and sex branches.

During single-task estimation, the output is the classification label corresponding to the most probable language. Due to the limited size of the dataset, the network is susceptible to overfitting. To address this issue, we employ dropout training (with a factor of 0.25) in the feedforward network.

4 Results and Analysis

The results and analysis section presents the findings of the experiments conducted. The three models were trained and evaluated in single and multi-task setups using the same data sets. The results are presented on the training, validation, and testing sets. The performance of the models was assessed using the accuracy metric formulated in equation 1.

Table 1 presents the results of the MobileNet base model. The first row of the table exhibits the accuracy achieved by the single-task model, while the second row contains the accuracy of the multi-task model trained using speaker identity as an auxiliary task. The training process of the single and multi-task models was stopped at epoch 20, as the accuracy reached a plateau at that value.

From Table 1 it can be seen that the language accuracy value on the test set for the multi-task architecture is higher than in the single-task. This could be indicative of the beneficial effect of incorporating speaker identity information into the proposed architecture, which enhances the model's discriminatory capacity in the SLR task.

It is also noteworthy that the discrepancy in language accuracy between the validation and testing sets diminishes as the single-task model transitions to the multi-task architecture, suggesting greater generalizability in the latter. In the case of Lozano's model adaptation, a second auxiliary task (speaker's sex) was included, and the same behavior was observed in Table 2.

This table demonstrates that multi-task outperforms single-task and that the gap between validation and test accuracy is smaller for the multi-task model. Table 3 presents the results of our frame-by-frame SLR approach. As anticipated, the performance of the SLR is degraded in comparison to the Mobile and Lozano models, given the 30-fold reduction in signal evaluation time (0.11s vs. 3s).

However, at the frame level, the speaker's identity and sex contribution to the SLR is well received by the multi-task model, which consistently improves performance. A fair comparison between the frame-level model and those based on three-second spectrograms can

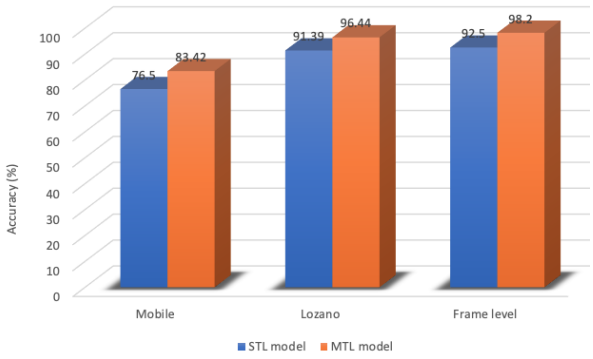

Fig. 6. SLR multi-task models' accuracy(%)

be made by deriving decisions about language identification at the utterance level by combining frame posteriors. A common and simple approach used in the literature is majority voting, where, at each frame, the language associated with the highest posterior receives a single vote while the rest receive none. The voting scheme aims to control the negative effect of outlier scores. The score for a given language l, s_l, is then computed by counting the received votes over all the N frames as follows:

$$s_l = \sum_{t=1}^{N} \delta(p(L_l|x_t, \theta)), \tag{3}$$

with δ function defined as:

$$\delta(p(L_l|x_t, \theta)) = \begin{cases} 1, & \text{if } l = \arg\max_l(p(L_l|x_t, \theta)), \\ 0, & \text{otherwise.} \end{cases}$$

where $p(L_l|x_t; \theta)$ represents the class probability output for the language l corresponding to the input example x_t at time t, by using the model defined by parameters θ.

Figure 5 collects the results of the majority voting over 1, 2, and 3-second intervals for the frame-level model. It is worth noting that for equal test utterance durations (3s), the frame-level approach performs better than the approaches of Mobile and Lozano. The posterior combination also demonstrates the improvement of the MTL approach in comparison to the STL.

The combination of language posteriors places frame-level model in a superior position in terms of performance, as shown in Figure 6. However, when considering the improvement of MTL, as shown in Table 4, the three models are quite similar, with a slight advantage for the MobileNet-based model.

5 Conclusions

All of the evaluations demonstrated the superiority of multitask learning for SLR when using speaker-related non-semantic characteristics, such as identity and sex, as auxiliary tasks. Initially, the hypothesis was verified for signals of three seconds, with and without image-based pretraining. Subsequently, an approach to deal with very short-duration utterances was proposed, and for both frame and utterance levels, MTL resulted in a superior SLR performance.

This article makes two main contributions: first, it determines the auxiliary tasks that should be used in a multi-task approach to SLR; second, it verifies that this approach will be beneficial. Additionally, it proposes a frame-level model with a convolutional neural network (CNN), which is a proposal much closer to real-time applications. The study and deepening of MTL for SLR offers possibilities for its use in low data resources environments.

References

1. **Abdullah, B. M., Avgustinova, T., Möbius, B., Klakow, D. (2020).** Cross-domain adaptation of spoken language identification for related languages: The curious case of slavic languages. Proceedings of the Interspeech 2020, pp. 477–481. DOI: 10.21437/Interspeech.2020-2930.
2. **Cai, W., Cai, D., Huang, S., Li, M. (2019).** Utterance-level end-to-end language identification using attention-based CNN-BLSTM. Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and

Signal Processing, pp. 5991–5995. DOI: 10.1109/ICASSP.2019.8682386.

3. **Caruana, R. (1993).** Multitask learning: a knowledge-based source of inductive bias. Proceedings of the Tenth International Conference on International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 41–48. DOI: https://doi.org/10.1016/B978-1-55860-307-3.50012-5.

4. **Chen, H., Xu, L., Yang, Z. (2018).** Multi-dimensional speaker information recognition with multi-task neural network. Proceedings of the IEEE 4th International Conference on Computer and Communications, pp. 2064–2068. DOI: 10.1109/CompComm.2018.8780705.

5. **Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., Ouellet, P. (2011).** Front-end factor analysis for speaker verification. Proceedings of the IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 4, pp. 788–798. DOI: 10.1109/TASL.2010.2064307.

6. **Furui, S. (1986).** Speaker-independent isolated word recognition based on emphasized spectral dynamics. Proceedings of the ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 11, pp. 1991–1994. DOI: 10.1109/ICASSP.1986.1168654.

7. **Gonzalez-Dominguez, J., Lopez-Moreno, I., Moreno, P. J., Gonzalez-Rodriguez, J. (2015).** Frame-by-frame language identification in short utterances using deep neural networks. Neural Networks, Vol. 64, pp. 49–58. DOI: 10.1016/j.neunet.2014.08.006.

8. **Kumar, R., Yeruva, V., Ganapathy, S. (2018).** On convolutional LSTM modeling for joint wake-word detection and text dependent speaker verification. Interspeech, pp. 1121–1125. DOI: 10.21437/Interspeech.2018-1759.

9. **Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J. (2019).** Direct modelling of speech emotion from raw speech. DOI: 10.48550/ARXIV.1904.03833.

10. **Li, Z., Zhao, M., Li, J., Zhi, Y., Li, L., Hong, Q. (2020).** The XMUSPEECH system for the AP19-OLR challenge. Proceedings of the Interspeech 2020, pp. 452–456. DOI: 10.21437/Interspeech.2020-1923.

11. **Lozano-Diez, A., Zazo-Candil, R., Gonzalez-Dominguez, J., Toledano, D. T., González-Rodríguez, J. (2015).** An end-to-end approach to language identification in short utterances using convolutional neural networks. Proceedings of the INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, ISCA, pp. 403–407. DOI: 10.21437/INTERSPEECH.2015-164.

12. **MacLean, K. (2009).** Voxforge. http://www.voxforge.org/.

13. **Mendes, C., Abad, A., Neto, J. P., Trancoso, I. (2019).** Recognition of latin american spanish using multi-task learning. Proceedings of the Interspeech 2019, pp. 2135–2139. DOI: 10.21437/Interspeech.2019-2772.

14. **Mohamed, A. (2014).** Deep neural network acoustic models for ASR. Ph.D. thesis, University of Toronto, Canada.

15. **Montalvo, A., Calvo, J. R., Bonastre, J. F. (2020).** Multi-task learning for voice related recognition tasks. Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, ISCA, pp. 2997–3001. DOI: 10.21437/INTERSPEECH.2020-1857.

16. **Montalvo-Bereau, A., Reyes-Díaz, F., Hernández-Sierra, G., Calvo-de-Lara, J. R. (2022).** Identificación de idioma hablado en señales cortas aplicando transferencia de aprendizaje. Revista Cubana de Ciencias Informáticas, Vol. 16, No. 1, pp. 77–91.

17. **Mukherjee, S., Shivam, N., Gangwal, A., Khaitan, L., Das, A. J. (2019).**

Spoken language recognition using CNN. Proceedings of the International Conference on Information Technology, pp. 37–41. DOI: 10.1109/ICIT48102.2019.00013.

18. **Padi, B., Ramoji, S., Yeruva, V., Kumar, S., Ganapathy, S. (2018).** The LEAP language recognition system for LRE 2017 challenge - improvements and error analysis. Proceedings of the Odyssey 2018: The Speaker and Language Recognition Workshop, pp. 31–38. DOI: 10.21437/ODYSSEY.2018-5.
19. **Palanisamy, K., Singhania, D., Yao, A. (2020).** Rethinking CNN models for audio classification. DOI: 10.48550/arXiv.2007.11154.
20. **Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. (2011).** The Kaldi speech recognition toolkit. Proceedings of the IEEE 2011 workshop on automatic speech recognition and understanding.
21. **Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., Chen, L. (2018).** MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474.
22. **Shen, P., Lu, X., Kawai, H. (2022).** Transducer-based language embedding for spoken language identification. Interspeech, pp. 3724–3728. DOI: 10.48550/ARXIV.2204.03888.
23. **Shon, S., Ali, A., Glass, J. R. (2018).** Convolutional neural networks and language embeddings for end-to-end dialect recognition. DOI: 10.48550/arXiv.1803.04567.
24. **Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., Quitry, F. d. C., Tagliasacchi, M., Shavitt, I., Emanuel, D., Haviv, Y. (2020).** Towards learning a universal non-semantic representation of speech. Interspeech 2020. DOI: 10.21437/interspeech.2020-1242.
25. **Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., Khudanpur, S. (2018).** Spoken language recognition using x-vectors. Odyssey 2018: The Speaker and Language Recognition Workshop, pp. 105–111. DOI: 10.21437/ODYSSEY.2018-15.
26. **Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., Khudanpur, S. (2016).** Deep neural network-based speaker embeddings for end-to-end speaker verification. Proceedings of the IEEE Spoken Language Technology Workshop, IEEE, pp. 165–170. DOI: 10.1109/SLT.2016.7846260.
27. **Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., Deller-Jr., J. R. (2002).** Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002), pp. 89–92. DOI: 10.21437/ICSLP.2002-74.
28. **Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., Courville, A. C. (2017).** Towards end-to-end speech recognition with deep convolutional neural networks. DOI: 10.48550/arXiv.1701.02720Focustolearnmore.
29. **Zhao, M., Li, R., Yan, S., Li, Z., Lu, H., Xia, S., Hong, Q., Li, L. (2019).** Phone-aware multi-task learning and length expanding for short-duration language recognition. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 433–437. DOI: 10.1109/APSIPAASC47483.2019.9023014.

Article received on 28/02/2024; accepted on 14/05/2024.
**Corresponding author is Ana Montalvo Bereau.*