Combining Deep and Unsupervised Features for Multilingual Speech Emotion Recognition

Vincenzo Scotti(✉), Federico Galati, Licia Sbattella, and Roberto Tedesco

DEIB, Politecnico di Milano, Via Golgi 42, 20133 Milan (MI), Italy
{vincenzo.scotti,licia.sbattella,roberto.tedesco}@polimi.it, federico.galati@mail.polimi.it

Abstract. In this paper we present a Convolutional Neural Network for multilingual emotion recognition from spoken sentences. The purpose of this work was to build a model capable of recognising emotions combining textual and acoustic information compatible with multiple languages. The model we derive has an end-to-end deep architecture, hence it takes raw text and audio data and uses convolutional layers to extract a hierarchy of classification features. Moreover, we show how the trained model achieves good performances in different languages thanks to the usage of multilingual unsupervised textual features. As an additional remark, it is worth to mention that our solution does not require text and audio to be word- or phoneme-aligned. The proposed model, PATHOSnet, was trained and evaluated on multiple corpora with different spoken languages (IEMOCAP, EmoFilm, SES and AESI). Before training, we tuned the hyper-parameters solely on the IEMOCAP corpus, which offers realistic audio recording and transcription of sentences with emotional content in English. The final model turned out to provide state-of-the-art performances on some of the selected data sets on the four considered emotions.

Keywords: Emotion recognition · Multilingual · Multi-modal analysis

1 Introduction

In the psychological literature, emotions are defined as a complex set of bidirectional interactions between physiological activation (arousal) and individual cognitive analysis (appraisal) [25]. This interaction generates affective experiences, cognitive processes and physiological adjustments, leading to the activation of adaptive behaviour [31]. In this regard, it's necessary to emphasise, as highlighted by many authors, the importance of the adaptive nature of emotions [11].

Partially supported by the European Union's Horizon 2020 project "WorkingAge" (grant agreement No. 826232).

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12662, pp. 114–128, 2021.
https://doi.org/10.1007/978-3-030-68790-8_10

Following this definition of emotions, literature highlights how an emotional state is characterised by some changes at the physiological level [21]. Such changes are an integral part of the emotion itself. Some physiological changes, such as acceleration of the heartbeat, increase in blood pressure, sweating, often occur without us being aware of them.

Specifically, the so-called *non-verbal communication* is a fundamental communication and expressive channel for emotions, as it is less consciously controllable. Examples of these non-verbal aspects are the *para-linguistic* ones, like voice tone, speech rate, pauses, silences, etc. On the other side, the *verbal communication*, with its *linguistic* aspects, is still a useful channel for the expression of emotions.

Different approaches, leveraging Neural Networks (NNs), have already shown that the combination of linguistic and para-linguistic clues can provide a useful contribution in the task of emotion recognition [2,6,36]. In such works, linguistic features have been mostly treated through pre-trained embedding models. Acoustic features, on the other hand, have been selected from pre-defined ones.

With this work, we were interested in two main aspects. The first one is understanding whether deep pre-trained features could lead to higher classification accuracy; hence, for a multi-modal analysis like this, pre-trained features for audio analysis should be used as well as pre-trained linguistic features. The second aspect is exploring the effects of a single model working with multiple languages at the same time; so, the multilingual model was built as an all-in-one model by feeding it with different corpora in different languages at train time.

The multilingual approach resulted in a training phase on a wider corpus which, in general, helps Deep Neural Networks (DNNs) to learn better features. Such features turned out to be correctly compatible with a multiple language environment as we expected. Notice that the multilingual approach allowed our model to deal with the data scarcity, which often prevents DNNs from being effectively trained.

Our classifier, called *Parallel, Audio-Textual, Hybrid Organisation for emotionS network* (PATHOSnet) reached an accuracy of 80.4% on the IEMOCAP [4] corpus (our main benchmark). The preceding best score of an automatic system, working with the same modalities, was obtained by Atmaja and colleagues: 75.5% [2]. Human listeners achieved 70% on the four emotions considered for this project, according to [6].

The rest of this paper is organised in the following sections. In Sect. 2 we present the state of the art for speech emotion recognition using NNs. In Sect. 3 we describe the data collections we use to train and test our model. In Sect. 4 we describe the input features used to feed our model. In Sect. 5 we describe the architecture of our model, for multilingual emotion recognition. In Sect. 6 we explain how we approached the training and evaluation processes. In Sect. 7 we report the results of the experiments and we comment on them. In Sect. 8 we sum up our work and provide hints about possible future work.

2 Related Works

In recent years speech emotion recognition has gained a lot of traction. Multimodal (audio, video and text) analysis has shown to be the correct way to address this problem. In particular, NN-based solutions have shown to produce better results. In fact, we're mainly interested in this kind of models for emotion recognition. Interested readers can refer to surveys [1], for other models.

For what concerns NN-based solutions, we noticed similar patterns in recent years for emotion recognition, where researchers started to employ, where possible, multimodal analysis on text and audio, and sometimes on video, too.

These input modalities are usually treated though pre-computed features. In particular, learnt semantic representations are used as linguistic features while handcrafted features are used for the acoustic part. The works we referred to are based on NNs and we considered as a main benchmark the IEMOCAP[4] corpus for emotion recognition (more on this in Sect. 3).

The best recent solution on the IEMOCAP corpus leveraged *Recurrent* or *Bi-Directional Recurrent* NNs (RNNs and BiRNN) [12,17,32], often including also an *attention mechanism* [3]. One of the first work on IEMOCAP with RNNs, however, reached only 54% classification accuracy [6], while a work proposing BiLSTMs with attention mechanism reached an accuracy of 71.0% through linguistic and acoustic analysis [36].

Even if useful to handle time series, the sequential structure of the RNNs makes their computations really slow; in fact, they cannot be parallelised [37]. Differently, Convolutional NNs (CNNs) are faster and easy to parallelise. For this reason, we implemented our NN using convolutional layers.

Authors of [24] proposed a deeper analysis of multimodal approaches for sentiment and emotion analysis. In their work, they focused on modality fusion and context usage. In particular, they found how the accuracy in the emotion recognition of a single utterance can be improved when leveraging information coming from the other utterances in the discourse. On IEMOCAP they reached an accuracy of 76.5% using audio-video and text, and an accuracy of 76.1% using solely audio and text. Even if this model produces impressive results, it relies on the usage of contextual information (i.e., a discourse), that might be not always available. Thus, we decided not to make use of such context information, working on individual, isolated sentences.

To our knowledge, none of the available NN models for emotion recognition applies a multilingual approach. Additionally, even if some of them are presented as deep learning approaches, they still employ manually-selected features, computed from the raw input, rather than resorting to deep models. With our work, we address both of these two aspects.

3 Corpora

Neural Networks are a data-driven framework; as such, they require labelled corpora to be trained on. For the purpose of this work, we considered different data sources in order to include different languages and use cases. In fact, it was our main interest to train a multilingual model, in order to provide a single tool available for everyone and cope with data scarcity. In particular, data scarcity represents a strong barrier for some languages.

To train our network, we selected the following corpora:

- Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [4];
- Emotional speech from Films corpus (EmoFilm) [29];
- Spanish Emotional Speech database (SES) [28];
- Athens Emotional States Inventory (AESI) [5].

In Table 1 it is possible to find statistics about such corpora, in terms of available samples, organised per-corpus and per-language.

The IEMOCAP corpus represents our main benchmark. It is the most complete and best managed of the considered corpora, but it only provides English samples. IEMOCAP employs both *categorical* [11] and a *dimensional* [13] representations of emotions in audio-visual data. The IEMOCAP corpus has been built recording ten actors in dyadic sessions, resulting in five sessions with two subjects each. Actors were asked to perform two tasks: play three selected scripts with clear emotional content, and improvise dialogues in hypothetical scenarios designed to elicit specific emotions. For the purpose of this work, and considering other works that used IEMOCAP, we selected only four basics emotions labels to discriminate among. The selected emotions are *Happiness*, *Anger*, *Sadness*, and *Neutral*.

The EmoFilm corpus contains samples in three different languages: English, Italian and Spanish. This corpus was built carefully, manually selecting audio recordings from a total of 43 movies. The search was conducted on the English dubs of the considered movies. Once the clips were identified, Italian and Spanish audio tracks were cut at the same time stamps. Rare emotional labels were excluded from the collection. As a result, 828 samples were retrieved (per-language) and labelled with the following emotions: Happiness, Anger, Sadness, *Fear* and *Contempt*. Fear was excluded from our work because there were enough samples of the same kind in other corpora; Contempt was discarded because it is generally not considered a basic emotion. No transcription was provided; for this reason, we resorted to an ASR[1] in order to have the textual content. This choice not only made us closer to real usage scenario, but also helped us to retrieve results that take into account possible transcription errors.

The SES corpus contains emotional speech recordings played by a professional male actor speaking Spanish. The available emotional labels in this corpus were Happiness, Anger, Sadness, Neural and *Surprsie*; the latter was excluded from our work because there were enough samples of the same kind in other corpora. The corpus is composed of several readings of the same neutral texts, displaying different emotions. On one side this aspect is useful as it will help to enforce the usage of acoustic features and prevent the model from sticking to a fixed vocabulary of words for the classification. On the other we had to ensure that

[1] https://gitlab.com/Jaco-Assistant/deepspeech-polyglot.

the model does not rely solely on acoustic features, hence we carefully analysed the results comparing the different corpora.

The AESI corpus is an audio-visual database for Greek emotion recognition. This corpus contains 696 recorded utterances in the Greek language by 20 native speakers. The emotional labels have been assessed through a survey. The samples are labelled according to one of these emotions: Happiness, Anger, Sadness, Neural and Fear. We included this corpus because we wanted to observe the generalisation capabilities of the network on smaller corpora when still provided with samples also in other languages.

IEMOCAP is the biggest of the considered corpora. As such, it was selected as our main benchmark and guided the choice of the emotional labels.

Table 1. Number of available samples (total and per-class) organised per-corpus and per-language

Corpus	Language	Number of samples				
		Emotion				Total
		Happiness	Anger	Sadness	Neutral	
IEMOCAP	English	1041	1103	1084	1708	4936
EmoFilm	English	70	77	74	0	221
	Italian	94	73	93	0	260
	Spanish	76	82	87	0	245
	All languages	240	232	254	0	726
SES	Spanish	732	725	728	1658	3843
AESI	Greek	139	139	140	139	557

4 Features

As premised we considered two distinct yet complementary input modalities for our network. On one side, we considered linguistic features, extracted from the transcription of the spoken sentence. On the other side, we considered the deep features extracted from the waveform of the spoken sentence. Both modalities use task-agnostic input features; in fact, none of the two modules generating features was specifically designed for emotion recognition. This was necessary since none of the data sets contains enough samples to train a deep model from scratch. Thus, in both cases, we retrieved pre-trained models, which we adapted for our work.

4.1 Linguistic Features

The *meaning* of what is uttered by the speaker, contained in transcriptions, represents an important piece of information for emotion recognition (i.e., the linguistic aspect). In fact, depending on the emotional context, certain words can be related to an emotional state more than others. To extract the textual features, we relied on word embeddings, a *vector semantics* representation [19] of words. Through word embeddings, every word is encoded as a vector in a d-dimensional space where words with similar meaning are encoded closely. Moreover, we trained the final network on multilingual embeddings, where words from different languages with the same meaning, are mapped in the same point in the embedding space. This was expected to help generalise across languages. In general, we expected that this semantically meaningful encoding will help the NN to associate similar meaning words to the same emotional state.

The embedded text is represented as a two-dimensional tensor, i.e. a matrix obtained embedding all the words in the utterance. The tensor is characterised by d columns, one for each dimension of the word embedding hyperspace, and a number of rows that matches that of the words in the sentence. The columns represent the sample's features, while the rows constitute the time axis.

During the hyperparameter-tuning phase, our model was fed with English-only embeddings. In particular, we used a *GloVe* model for word embeddings [30] (with 300-dimensional vectors). As premised, we used a pre-trained model[2]. Subsequently, the final model was trained with multilingual embeddings, by means of *MUSE* framework [8,20]. These multilingual embeddings are obtained starting from pre-trained *FastText* word embedding models [26] in different languages (always with 300-dimensional vectors). The embeddings are then transformed so that corresponding words in the different languages result in overlapping vectors. As for the English model, we used pre-trained MUSE embeddings[3].

4.2 Acoustic Features

We used acoustic features to capture the information about *how* a person is talking (i.e., the para-linguistic aspect). The choice of deep acoustic features represents a strong change with respect to previous work in emotion recognition. In fact, to our knowledge, previous works relied solely on pre-defined acoustic features [6]. Such features were manually selected to highlight the aspects of the voice signal that were expected to correlate the most and to cope with the reduced amount of samples. We decided, instead, to use a transfer learning approach [38] and rely on the features extracted by deep models trained on huge classification tasks.

The DNNs we employed to extract features were designed and trained using the same concepts and huge audio classification data sets. In fact, they employed the same architectures of image recognition NNs, adapted to take as input the (mel-filtered) spectrogram of the vocal signal. Their basic idea is to threat the spectrogram as an image and use 2-D CNNs to learn a feature hierarchy useful for audio classification. Thanks to the use of a huge audio data set, the models were able to produce very general features, which resulted to be transferrable

[2] https://nlp.stanford.edu/projects/glove/.

[3] https://github.com/facebookresearch/MUSE.

across different tasks. This is the same approach used with image recognition models trained on *ImageNet* [9]: the CNNs are trained as classifiers, then their classification heads are removed to transfer the features to other image analysis tasks.

We experimented with two different networks. The former was *VGGish* [16], a variant of *VGG* [33], which is a NN for image recognition. VGGish was trained on the *AudioSet*[4] corpus [14]. The feature extraction variant takes as input a 64 bin log-scaled, mel-filtered spectrogram (computed with a window size of 25 ms and a hop size of 10 ms) and produces a 128-dimensional feature vector for every non-overlapping 0.96 s window in the input. The latter network was *Thin ResNet-34* with *GhostVLAD* pooling layer [41], a variant of *ResNet* [15], a NN for image recognition. This second feature extraction network takes as input a 257 bin normalised spectrogram (computed with a window size of 25 ms and a hop size of 16 ms) and produces a 512-dimensional feature vector for every non-overlapping 0.045 s window in the input. We relied on a Keras[5] implementation for both VGGish[6] and Thin ResNet-34 with GhostVLAD[7].

The acoustic features are then represented similarly to the linguistic ones: they are managed as a two-dimensional tensor. The row axis represents the time dimension, the column axis represents the features (conceptually this is similar to a spectrogram with its bins).

5 Model

The model we developed, called PATHOSnet and represented in Fig. 1, is a multi-modal DNN for emotion recognition built upon transferred deep features. The model is composed of two parallel branches, one for linguistic analysis and one for acoustic analysis, which are later merged together. A high-level view of this model is depicted in Sect. 5. These two symmetric branches are CNNs, composed of 1-D convolutional and pooling layers. A depiction of such blocks is reported in Sect. 5. We inserted each of the two blocks on top of the corresponding feature extractor. The classifier on top is, instead, a fully-connected layer with a softmax activation function over the four considered classes.

The convolutional blocks in the two branches are designed like those of a ResNet [15] network, adapted for the 1-D scenario. To flatten the information along the time axis and produce a single feature vector for each modality, we relied on a Global Average Pooling (GAP) layer [22]; GAP not only allows to "compress" spatial information, averaging along the time axis, but it does so with a low computational effort (differently from *attention*-based solutions [27]).

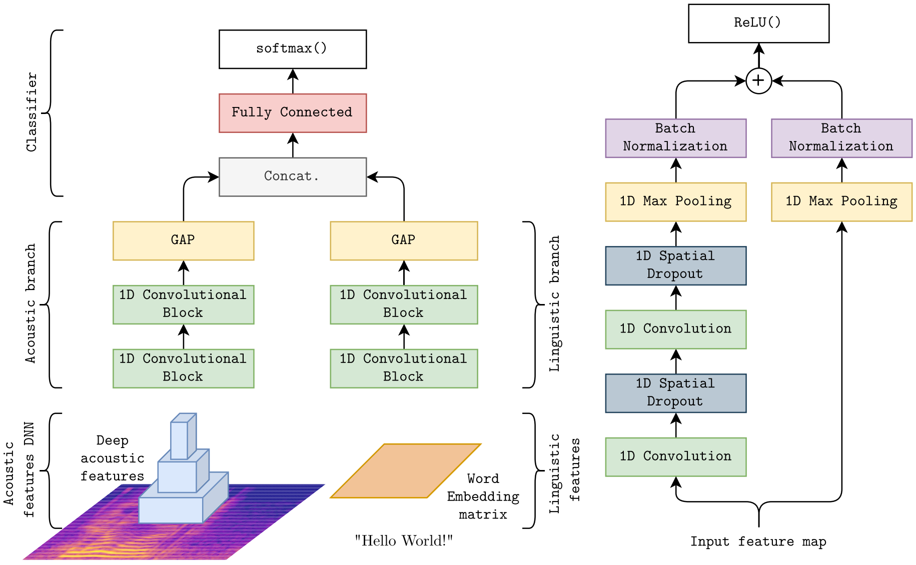
To merge the two branches of the network we adopted a simple feature fusion approach [1]: we concatenated the two feature vectors coming from the two separate branches and learnt a fully connected transformation to combine the vectors

[4] https://research.google.com/audioset/.

[5] https://keras.io.

[6] https://github.com/beasteers/VGGish.

[7] https://github.com/taylorlu/ghostvlad-speaker.



(a) High level model view.

(b) Convolutional block detail.

Fig. 1. PATHOSnet internal structure.

directly into the class probabilities. In this way, acoustic features are not required to be aligned with the textual ones. In fact, each branch takes care of embedding the "temporal" information in its intermediate representation, removing the time axis by the end of its transformation. This was a great advantage as this kind of alignment either isn't available or is difficult to obtain.

In order to enforce regularisation and avoid overfitting we adopted the *spatial dropout* [35] and the *batch normalisation* [18]. Regularisation was also enforced thanks to the GAP layer. Finally, to avoid overfitting we employed the *early stopping* approach.

As an additional note, we want to point out that the DNN used to extract the acoustic features is an integral part of PATHOSnet; in this way, we managed to perform fine-tuning of its weights. In Sect. 6 we provided more details about this part.

PATHOSnet is an example of so-called *ensemble* models. In our case, this approach was useful as the textual-based and acoustic-based networks turned out to be complementary. To build the ensemble, we removed the classifier on top of the two networks, and learnt a new linear classification function on top of the concatenated feature vectors.

The implementation of our model was realised through the Keras framework, using Tensorflow[8] as backend. The entire code, from feature extraction to the NN was developed solely through the Python programming language.

[8] https://www.tensorflow.org.

6 Experiments

In our experiments, we followed the classic train-test steps adopted by NN frameworks. The training process of the networks was divided into 2 steps:

1. train the convolutional blocks above a single couple of DNN for acoustic features and word embeddings model. In this first phase the weights of the acoustic DNN were “frozen”;
2. fine-tune the single networks “unlocking” the first layer of the DNN for acoustic features (we experimented unlocking more layers but without good results).

Testing was conducted on the same percentages of data from all the considered corpora and languages.

We trained three separate versions of PATHOSnet:

1. We trained an English-only model on IEMOCAP. We used only VGGish deep features for the acoustic part. We used GloVe word embeddings for the linguistic part. We referred to this model as *baseline*;
2. We trained an English-only, ensemble model on IEMOCAP. We used both VGGish and Thin ResNet-34 with GhostVLAD deep features for the acoustic part. We used multilingual MUSE word embeddings for the linguistic part; We referred to this model ensemble as *PATHOSnet*;
3. We trained the multilingual ensemble model on all corpora. We used both VGGish and Thin ResNet-34 with GhostVLAD deep features for the acoustic part. We used multilingual MUSE word embeddings for the linguistic part. We referred to this model ensemble as *PATHOSnet (multilingual)*;

In order to provide more robust results, we resorted to 5-fold cross-validation. In this way, a fifth of each corpus was used as the test set in all experiments. Additionally, we further split (always corpus-wise) the training data into train and validation sets (enabling to use early stopping). We used an 80–20% train-validation split to further separate validation samples. Splitting was done randomly but taking into account corpus and class sizes.

Before training the network on the multilingual corpora we performed hyper-parameter tuning. This tuning was performed only on the IEMOCAP corpus, and on a single model using English word embeddings and VGGish features.

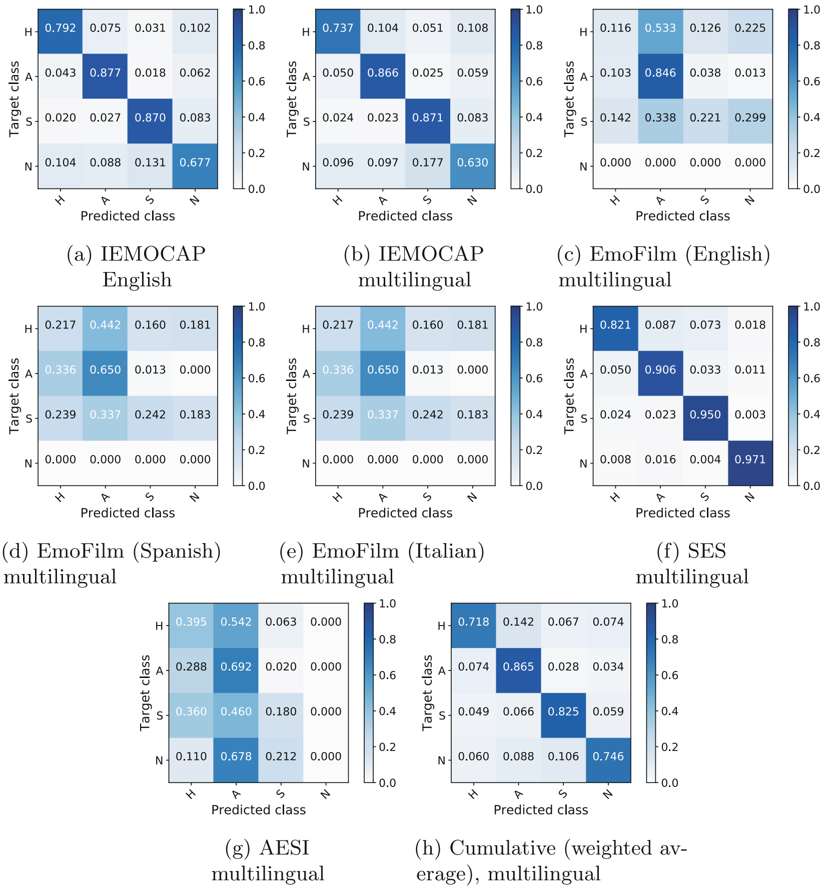
The derived hyper-parameters are the following. We selected the *RMSprop* [34] optimiser, with an initial learning rate $l_0 = 0.001$; the learning rate was decayed exponentially at each epoch e down to a minimum of 0.00001 using the function in Eq. (1).

$$l(e) = \max(l_0 \cdot \exp(0.1 \cdot e), 0.00001) \tag{1}$$

During fine-tuning phases, the learning rate restarted from that of the last epoch. The kernels in the convolutional blocks covered all a width of 3 time steps, the linguistic branch used a feature size of 128, while the acoustic branch used a feature size of 256. Each branch used two convolutional blocks. The network was trained for 45 epochs at each step, using early stopping.

7 Results

The main results are reported in Table 2. We reported the values from the main classification metrics. As premised these values are obtained through cross-validation, the scores of each fold are aggregated through a weighted mean on the number of samples per class. In Fig. 2 are reported, instead, the confusion matrices of the English ensemble model and the multilingual ensemble model (for the latter we reported cumulative results as well as results divided across corpora).



(a) IEMOCAP English

(b) IEMOCAP multilingual

(c) EmoFilm (English) multilingual

(d) EmoFilm (Spanish) multilingual

(e) EmoFilm (Italian) multilingual

(f) SES multilingual

(g) AESI multilingual

(h) Cumulative (weighted average), multilingual

Fig. 2. Confusion matrices computed on the test sets of the considered corpora and languages, plus the combined results. Values are averaged among the five folds. Legend: Happiness (H), Anger (A), Sadness (S), Neutral (N).

The first remark we point out is that even the baseline model outperforms the state of the art. This highlights how deep features transferred from another

Table 2. Classification results of the proposed model. The reported baseline is from the hyperparameters tuning phase (VGGish and GloVe features). All the reported values from PATHOSnet are computed using the deep features ensemble model. All the reported values are computed through a weighted average on the support of each class. The metrics are accuracy (Acc.), precision (Prec.), recall (Rec.), F1-score (F1) and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC).

Model	Language	Corpus	Metric (%)				
			Acc.	Prec.	Rec.	F1	AUC
Humans [6]	English	IEMOCAP	70.0	–	–	–	–
Atmaja's (previous state of the art)	English	IEMOCAP	75.5	–	–	–	–
Baseline	English	IEMOCAP	77.0	75.7	74.1	73.4	93.4
PATHOSnet	English	IEMOCAP	80.4	79.1	78.8	78.6	94.6
PATHOSnet (multilingual)	English	IEMOCAP	77.6	76.4	75.8	75.4	93.4
		EmoFilm	39.4	47.6	45.5	40.7	70.1
	Spanish	EmoFilm	37.0	46.1	41.4	39.6	65.9
		SES	91.2	92.8	92.7	92.7	99.2
	Italian	EmoFilm	37.7	38.8	37.7	34.3	67.8
	Greek	AESI	31.7	25.6	31.7	24.6	62.9
	Cumulative		78.5	77.2	77.0	78.5	92.6

task are more useful than those manually selected. Moreover, with respect to Atmaja's work, we produced a way smaller network. In total this version of PATHOSnet has a similar number of parameters (around 5 million) but only slightly more of 1 million of them are trainable in our case (in Atmaja's work they were all trainable). These parameters are those of the convolutional blocks and classification layer. The remaining parameters in PATHOSnet come from the lower layers of VGGish, which we integrated into our network. Moreover, we haven't used any LSTM or attention mechanism but only convolutions, pooling and a single dense layer. This underlined once again how the choice of correct features is crucial to obtain better results.

For what concerns the ensemble model, it outperformed the baseline reaching a weighted mean test accuracy of 80.4%. The score is weighed taking into account samples in each data set and for each language of the data set. Judging from the confusion matrix in Sect. 7, the usage of two models for deep acoustic features and the FastText multilingual embeddings helped to better separate all classes. From the confusion matrix, we see that the hardest class to separate is the Neutral one; Anger and sadness are instead the easiest classes to separate. To our knowledge, these are the best results ever obtained on the IEMOCAP corpus.

Finally, the ensemble for multilingual emotion recognition obtained on average, across languages, competitive results: the accuracy was higher than 78% and all the other metrics confirm the goodness of the model (AUC is over 0.9). However, the analysis on single languages, showed that in some cases the network didn't meet our expectations. Still, from the single languages scores, we saw that the model is still behaving better than random guessing. This can be seen by the fact that the *AUC* is always higher than 0.5.

On sufficiently big data sets, the model still shows impressive generalization capabilities across languages. This can be seen by the performances on IEMOCAP for English (still better than the previous state of the art and our baseline). On SES for Spanish, the same applies, the model achieves an accuracy even higher of that on IEMOCAP.

On smaller data sets (EmoFilm and AESI) the multilingual network showed lower scores. We believe that the lower results on EmoFilm are due to transcription errors, introduced by the ASR. The model showed similar results in terms of accuracy among the three languages of this data set. It is also clear that Italian suffers from the lack of other samples with respect to English and Spanish, judging by the fact that it has lower scores among the three. This, however, shows again how linguistic features are important and strictly rely on correct transcriptions. Finally, as for the Italian part of EmoFilm, AESI most probably showed lower results because of the data scarcity. Interestingly, from the confusion matrix in Sect. 7, we see that on the Greek language Neutral label is never predicted and is most often misclassified as anger.

8 Conclusions

In this paper we presented the architecture of a multi-modal NN for multilingual speech emotion recognition, which leverages linguistic and acoustic features. Multilingual word embeddings was used to generate the linguistic features needed by the network working on text. The network working on voice is trained through transfer learning and fine-tuning, on top of different, pre-trained networks that generate acoustic features. We then merged these two networks into an ensemble model, to achieve a better classification accuracy. The model we presented achieved the state of the art classification accuracy on different emotion recognition corpora. Moreover, we trained a single model with different languages, showing how it is possible to take multiple languages into account at the same time.

The experiments we performed partly confirmed our hypotheses. Deep unsupervised acoustic features are better than hand-crafted for emotion recognition. Results on IEMOCAP and SES confirmed also that it is possible to obtain a multilingual model, provided sufficiently big corpora for all languages. On this same side, results on Italian EmoFilm and AESI showed that languages with poor data sets are harder to integrate. Finally, results on EmoFilm showed how badly the errors introduced by the ASR influence the recognition capabilities.

The first step we are going to do is investigating the source of the errors in order to obtain acceptable results in all languages. In the future, we are planning to improve and extend our model under multiple aspects. Even though "vanilla" word embeddings seems to provide an efficient representation, we are interested in observing the results using *contextual embeddings* [23]. Such representation turned out to be very informative for many tasks [39,40], hence we can leverage some the multilingual transformer models like BERT [10] or XLM [7].

At the same time, we are also interested in extending the set of languages we consider. Both the employed word embedding models and the suggested contextual embedding ones already support more languages than the ones we used, hence what we will require are labelled corpora is such languages.

Finally, we plan to extend the emotion the model is able to handle, including at least the six basic ones identified by Ekman [11]. Alternatively, we could resort to continuous representations [13]; however, it would require feasible labelled corpora for all the language we consider.

References

1. Akçay, M.B., Oğuz, K.: Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. **116**, 56–76 (2020)
2. Atmaja, B.T., Shirai, K., Akagi, M.: Speech emotion recognition using speech feature and word embedding. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 519–523 (2019)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2016)
4. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335 (2008)
5. Chaspari, T., Soldatos, C., Maragos, P.: The development of the Athens emotional states inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences. World J. Biol. Psychiatry **16**(5), 312–322 (2015)
6. Chernykh, V., Sterling, G., Prihodko, P.: Emotion recognition from speech with recurrent neural networks. CoRR abs/1701.08071 (2017)
7. Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems, pp. 7059–7069 (2019)
8. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data (2018)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR09 (2009)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
11. Ekman, P.: An argument for basic emotions. Cogn. Emotion **6**, 169–200 (1992)
12. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**(2), 179–211 (1990)
13. Fridlund, A.J.: Human Facial Expression: An Evolutionary View. Academic Press, San Diego (1994)
14. Gemmeke, J.F., et al.: Audio set: an ontology and human-labeled dataset for audio events. In: Proceedings of IEEE ICASSP 2017, New Orleans, LA (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)

16. Hershey, S., et al.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135, March 2017. https://doi.org/10.1109/ICASSP.2017.7952132
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–80 (1997)
18. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift (2015)
19. Jurafsky, D., Martin, J.H.: Speech and Language Processing, chap. 6: Vector Semantics and Embeddings. Prentice-Hall, 3rd edn., August 2020, draft of August 2020. https://web.stanford.edu/~jurafsky/slp3/
20. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only (2018)
21. Lazarus, R.S.: Emotion and adaptation. Oxford University Press on Demand 1, 35–54, May 1991
22. Lin, M., Chen, Q., Yan, S.: Network in network (2014)
23. Liu, Q., Kusner, M.J., Blunsom, P.: A survey on contextual embeddings (2020)
24. Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., Poria, S.: Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowl. Based Syst. **161**, 124–133 (2018)
25. Manstead, A.S., Wagner, H.L.: Arousal, cognition and emotion: an appraisal of two-factor theory. Cogn. Emotion **1**, 35–54 (1992)
26. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)
27. Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231, March 2017
28. Montero, J.M., Gutiérrez-Arriola, J., Colás, J., Macías-Guarasa, J., Enríquez, E., Pardo, J.M.: Development of an emotional speech synthesiser in Spanish. In: Sixth European Conference on Speech Communication and Technology (1999)
29. Parada-Cabaleiro, E., Costantini, G., Batliner, A., Baird, A., Schuller, B.W.: Categorical vs dimensional perception of Italian emotional speech. In: INTERSPEECH, pp. 3638–3642 (2018)
30. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014). http://www.aclweb.org/anthology/D14-1162
31. Plutchik, R.: The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. **89**(4), 344–350 (2001)
32. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. Trans. Sig. Proc. **45**(11), 2673–2681 (1997)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
34. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA Neural Netw. Machine Learn. **4**(2), 26–31 (2012)
35. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks (2015)

36. Tripathi, S., Tripathi, S., Beigi, H.: Multi-modal emotion recognition on IEMOCAP dataset using deep learning (2019)
37. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
38. Ventura, D., Warnick, S.: A theoretical foundation for inductive transfer. Brigham Young University, College of Physical and Mathematical Sciences (2007)
39. Wang, A., et al.: SuperGLUE: a stickier benchmark for general-purpose language understanding systems (2020)
40. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: a multitask benchmark and analysis platform for natural language understanding (2019)
41. Xie, W., Nagrani, A., Chung, J.S., Zisserman, A.: Utterance-level aggregation for speaker recognition in the wild. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5791–5795. IEEE (2019)