In: Proc. of the IEEE International Joint Conference on Neural Networks (IJCNN/WCCI), Vancouver, Canada (July, 2016)

Learning Auditory Neural Representations for Emotion Recognition

Pablo Barros
Department of Computer Science
University of Hamburg
Hamburg, Germany
barros@informatik.uni-hamburg.de

Cornelius Weber
Department of Computer Science
University of Hamburg
Hamburg, Germany
weber@informatik.uni-hamburg.de

Stefan Wermter
Department of Computer Science
University of Hamburg
Hamburg, Germany
wermter@informatik.uni-hamburg.de

***Abstract*—Auditory emotion recognition has become a very important topic in recent years. However, still after the development of some architectures and frameworks, generalization is a big problem. Our model examines the capability of deep neural networks to learn specific features for different kinds of auditory emotion recognition: speech and music-based recognition. We propose the use of a cross-channel architecture to improve the generalization aspects of complex auditory recognition by the integration of previously learned knowledge of specific representation into a high-level auditory descriptor. We evaluate our models using the SAVEE dataset, the GTZAN dataset and the EmotiW corpus, and show comparable results with state-of-the-art approaches.**

I. Introduction

One of the pillars of communication between humans is the capability to perceive, understand and respond to social interactions, usually determined through affective expressions [1]. Understanding emotion expressions can improve the processing and reaction of automatic emotion recognition systems, such as robots or expert systems, to natural human behavior, which improves the efficiency and complexity of human-machine interaction [2]. If a robot can identify emotional expressions of humans, it can alter its interaction with the environment [3]. It can improve its own capability of solving problems by using these expressions as part of its own decision making process [4].

Emotional expression recognition has been shown to be a very difficult task and attracted a lot of research in recent years. There is no consensus in the literature to define emotions [5], but the observation of several characteristics, and among them auditory information, are used in their identification. Auditory recognition is complex since not only human speech carries emotional information, but music and ambient sounds also. Each one of these categories carries its own information, which usually is represented by different characteristics. In speech, it is possible to measure affect by evaluating fundamental frequency and loudness [6], or voice quality [7] for example, and when classifying music, characteristics such as tempo and instrumentation [8] are usually used.

Each auditory feature carries its own information, changing the nature of the audio representation. For the same clip of sounds, very distinct information can be extracted for different tasks. Thus, Aljanaki et al. [9] use a set of three different descriptors, namely chroma features, loudness and Mel-Frequency Cepstral Coefficients (MFCC), to represent distinct auditory information. They also obtain different temporal/non-temporal representations for each descriptor, using sliding windows for discrete representations or Hidden Markov Models for temporal dependencies. A total of 83 feature representations are obtained. After the feature extraction, they use a hierarchical non-parametric Bayesian model with a Gaussian process to classify the features. They show that their approach has a certain degree of generalization, but the exhaustive search for tuning each parameter of the model for multiple feature representations and feature combinations is not a viable option.

Similar to the work of [9], several approaches [10], [11], [12] use an extensive feature extraction strategy to represent the audio input: they extract several features, creating an over-sized representation to be classified. The strength of this strategy relies on redundancy. The problem is that usually it is not clear how well each of these descriptors actually represents the data, which can lead to not capturing the essential aspects of the audio information, model overfitting and decreasing the generalization capability [13].

Recently the development of deep learning for audio representation and classification, especially speech, showed how to increase generalization by using a learned audio representation instead of a huge set of audio descriptors [14]. In this strategy, the network is able to learn how to represent the auditory information in the most efficient way for the task. Initial research was applied to music [15] and speech recognition [16], and was shown to be successful in avoiding overfitting. The problem with this approach is that it needs an extensive amount of data to learn high-level audio representations, especially when applied to natural sounds, which can include speech, music, ambient sounds and sound effects [17].

Based on the initial success of deep learning strategies, we propose a Convolutional Neural Network (CNN) approach for auditory emotion recognition. Our model develops a different training strategy named cross-channel learning, which was applied previously to emotion recognition in the visual modality. Using a two-channel CNN, we now extend our model to learn specific auditory representations for two different audio

modalities: music and speech. Each of the channels is trained individually with specific data, and a cross-channel is trained to fuse the specific representations of each channel into a higher level audio descriptor.

To evaluate our model, we use three different datasets. The SAVEE dataset is used to train and evaluate the Speech-specific architecture, the GTZAN dataset for the Music-specific architecture, and the Emotions-in-the-Wild-Challenge (EmotiW) dataset is used to train the Cross-channel architecture and evaluate the high-level audio representation. The EmotiW dataset contains clips of movies with natural emotion expressions with speech, music, ambient sounds and sound effects. The use of such a challenging dataset allows us to evaluate our model for the generalization of the learned features, and its application in natural emotion recognition tasks. After series of experiments are performed, we show that our Cross-channel architecture yields compatible results with state-of-the-art approaches on the EmotiW dataset, and is able to learn general features from speech-specific data and music-specific data.

We first introduce our model and the training strategy in Section II. Section III describes the experiments methodology and presents the results. In Section IV we discuss our results and the network behavior for different input. Finally, in Section V we present our conclusions and future work.

II. Learning auditory representations

Our proposed model is a Convolutional Neural Network (CNN)-based approach, which receives an audio stream as input and classifies it into different emotional expression classes. Our network implements a Multichannel Convolutional Neural Network (MCCNN) [18], where each channel of the network is driven to learn specific information from the input signal. One of the channels is trained with only speech information and the other only with music. Each channel requires a different pre-processing of the input signal, and generates its own high level audio representation.

We train our network in two steps: first each of the channels is trained individually to learn how to represent specific data structures. The second step uses the learned filters in a MCCNN architecture, where both channels are connected to a Cross-Channel layer [19]. During the second training step, only the Cross-Channel is trained and it learns how to integrate the already learned high-level audio representation of each channel.

We use two different architectures, one Music-specific and one Speech-specific, to train each of the channels. Each architecture is trained with a different dataset, with music specific data, the GTZAN corpus, and with speech specific data, the SAVEE corpus. This allows us to enforce that each architecture will learn very specific features from the data, specializing the filters in describing the audio for speech and music modalities. This ensures that when training our Cross-channel architecture, it will learn only how to correlate these two specific representations, and thus improving the generalization of the model.

A. Convolutional Neural Networks

CNNs have been applied with success to several classification tasks [20], [21], usually related to image recognition. CNNs are composed of a stack of layers, each one containing usually two operations: convolution and pooling. These operations simulate the response of simple and complex cell layers in the brain's visual cortex V1 region [22].

The convolution operation simulates the simple cell responses, and uses local filters to compute high-order features from the input signal. After each convolution, a pooling operation is applied which represents the responses of complex cells. This increases the invariance to geometric distortion by pooling the convolution units that belong to the same receptive field in the previous layer.

To increase the number of features which are extracted by simple cells, each convolution layer has a series of different filters which are applied to the input signal. This operation generates a different set of features per filter, creating a feature map. The complex cells pool units from the receptive fields in each of the feature maps, decreasing the dimensionality of the data and increasing the complexity of abstraction of each feature map. After a series of layers containing convolution and pooling operations, the input signal has a substantially reduced dimensionality, but a very high level of representation.

The activation of each convolution unit v_{nc}^{xy} at (x,y) of the nth filter in the cth layer is given by

$$v_{nc}^{xy} = max\left(b_c + \sum_m \sum_{h=1}^{H} \sum_{w=1}^{W} w_{(c-1)m}^{hw} v_{(c-1)m}^{(x+h)(y+w)}, 0\right), \tag{1}$$

where $\max(\cdot, 0)$ represents the rectified linear function, which was shown to be more suitable for training deep neural architectures, as discussed by [23], b_c is the bias for the current feature map of the cth layer, m indexes over the set of feature maps in the (c-1) layer connected to the current layer c, $w_{(c-1)m}^{hw}$ is the weight of the connection between the unit (h,w) within a receptive field, connected to the previous layer, $c-1$, and to the filter map m. H and W are the height and width of the receptive field.

A receptive field of simple cells is connected to a complex cell, and a pooling operation is applied. Max-pooling is usually performed, where each complex cell outputs the maximum activation of the receptive field $u(x, y)$ and is defined as:

$$a_j = max\left(v_c u(x, y)\right), \tag{2}$$

where v_c is the output of the simple cell. In this function, the complex cell computes the maximum activation among the receptive field $u(x, y)$. The maximum operation down-samples the feature map, maintaining the simple cell structure.

The parameters of a CNN could be learned either by a supervised approach tuning the filters in a training database, as presented by [24], or an unsupervised approach as present in [25]. Our proposed model uses the supervised approach. Even for approaches where unsupervised training is used, such

as deep belief networks or stacked-auto encoders, the use of supervised fine tuning is advised [26]. The use of supervised training allows us to train the network with a smaller amount of data, compared to when using unsupervised training.

B. Convolutional Neural Networks for audio signals

Audio representations vary depending on the nature of the input. There is evidence that shows that the use of Mel-Cepstran Coefficients (MFCC) is suited for speech signals [14], but does not provide much information when describing music. On the other hand, spectral representations such as power spectrograms, lead to good results on music classification tasks [27]. Most research on auditory emotion recognition just uses several audio descriptors, not taking into consideration the different nature of the audio input.

In our work, we represent the audio signal in two different ways: for the speech-specific architecture we use MFFCs and for the Music-specific architecture we use a power spectrogram.

To obtain the power spectrogram, the audio is separated into smaller clips, and a Discrete Fourier Transform (DFT) is applied to each clip, transforming the audio into the frequency domain. The power spectrogram describes the distribution of frequency components on each clip.

MFCCs are coefficients derived from a cepstral representation of an audio sequence. First, the local power spectrum of a part of the input sound is created, based on a linear cosine transform. Then, the spectrum is transformed on the Mel scale of frequency. The mel scale is closer to the human auditory system's response than the linear frequency, used by the common spectrogram representation.

Convolutional Neural Networks were used to acoustic modeling [28], [29], [30]. There is a difference when training CNNs with audio signals and with images. Usually, CNN inputs are 2D images which make the filters into a 2D matrix applied to both hight and width of the image. The filter is applied to different parts of the image, and each filter captures local information of the region where it is applied.

To be used in a CNN, the audio signal is usually represented as a 2D image in the form of a spectrogram. However, in a spectrogram each axis represents different information, where the X axis represents time and the Y axis the spectral representation. For the power spectrogram representation every value along Y is obtained by a clip of the audio which is directly represented by the value of X. So, the topological processing of the 2D filters actually learns local representation of adjacent features.

On the MFCCs, the use of 2D filters has been shown to be a problem. As described by Abdel-Hamid et al. [30], because of the cosine transform, each value of Y is projected in the mel frequency basis, which may not maintain locality. This way, because of the topological nature of the 2D filters, a filter will try to learn patterns in adjacent regions, which could be non-existent. To solve this problem, Adbel-Hamid et al. propose the use of 1D filters, instead of the common 2D filters. The convolution process is the same, but the network learns how to represent the data for each part of the audio, and not look at the local information shared by local neighbors. The pooling is applied in each filter map separately and at the end the network learns to represent different coefficients and not the interconnection among them. This strategy proved to be successful and it is used in our speech-specific architecture.

C. Music-specific architecture

We feed the network with clips of one second of audio input which are re-sampled to 16000 Hz. For the Music-specific architecture, we compute the power spectrum over a window of 25ms with a slide of 10ms. The frequency resolution, the number of bins used to compute the Fourier Transform, is 2048. This generates a spectrogram with a dimension size 1024 x 136, which means 1024 features, each one with 136 descriptors. To reduce the computational cost to train our network, we resize the image by a factor of 8, resulting in an input size of 128 x 17. During our experiments, we found that reducing the size by a factor of 8 improved the training and recognition time, and did not reduce significantly the classification performance of the network.

The empirically determined architecture has two layers, each of them performing convolution and pooling, attached to a flattening hidden layer which is connected to a logistic regression layer. The first layer contains 10 filters, each one with a dimension of 5x5 and applies the pooling in a receptive field with dimension 2x2. The second layer has 20 filters, with dimension 3x3 and pooling over a receptive field of size 2x2. The hidden layer has 500 units, and all the layers implement rectified linear units. The network is trained with backpropagation and Figure 1 illustrates the Music-specific architecture.

D. Speech-specific architecture

The MFCCs are computed on one second clips which are re-sampled to 16000Hz, the same way as we did with the Music-Architecture. To conpute the MFCCs we use a window of 25ms with a slide of 10ms and a frequency resolution of 1024. Based on our experimental results, we decided to choose the first 26 coefficients for each segment, which resulted in a spectrogram with dimensions of 35x26, meaning 32 bins, each one with 26 coefficients.

On this architecture we apply the with 1 dimension filters [30]. The network has three layers which are fully connected to a hidden layer, which is then attached to a logistic regression layer. The first layer has 5 filters, each one with a dimension 1x3, and applies a pooling over a receptive field with dimension 1x2. The second layer has 10 filters, with the same dimensions and applies the pooling over the same receptive field size. The third and last convolution layer has 20 filters with dimension of 1x2 and pooling over a receptive field of 1x2.

In this network, only the coefficients are learned and pooled, resulting in a high-level representation which keeps the same temporal topology of the audio signal. This network has 500 units in the hidden layer, implements rectified linear units and

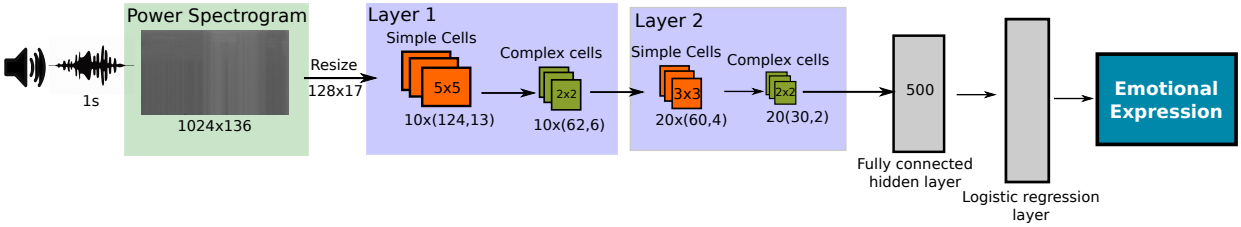

Fig. 1. CNN architecture for Music-specific representation. We show 1s of audio in a power spectrogram form before sending it to the network. The network has two layers, the first one with 10 filters with dimension 5x5 and pooling 2x2. The second has 20 filters with dimension 3x3 and we apply a pooling over a 2x2 receptive field. The fully connected hidden layer has 500 units.

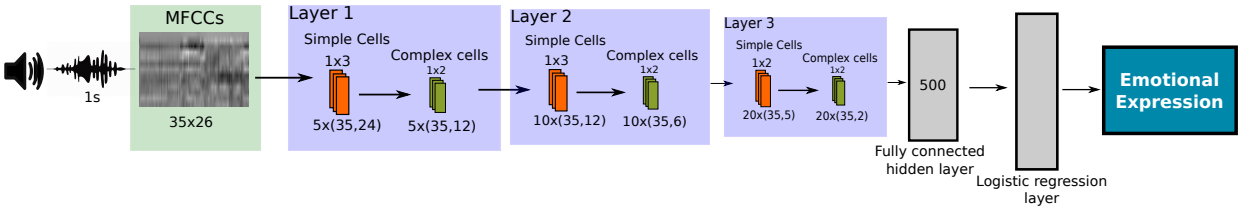

Fig. 2. CNN architecture for Speech-specific representation. We extract the MFCCs of 1s of audio before sending them to the network. The network has three layers, the first one with 5 filters with dimension 1x3 and pooling 1x2, meaning 1D convolution and pooling. The second has 10 filters with dimension 1x3 and applies a pooling over a 1x2 receptive field. The last layer has 20 filters with size 1x2 and pooling 1x2. The fully connected hidden layer has 500 units.

is trained with backpropagation and Figure 2 illustrates the Speech-specific architecture.

E. Cross-channel learning strategy

Our Cross-Channel Convolutional Neural Network was developed previously for visual emotion recognition tasks [19]. The Cross-channel is an extra convolution channel which receives as input a very high level representation of the network's input data. Different from the visual task, where the cross-channel received representations of face expression and movement, here we propose the Cross-channel receiving features with two different representations from the same audio signal.

We build an MCCNN architecture with two channels, each one containing the convolution filters learned by the Music-and-Speech-specific architectures, without using the fully-connected hidden layer nor the logistic regression layer. We see this approach as a pre-training strategy, where we train each channel individually to learn how to represent specific features from the same audio input, and the Cross-channel to integrate them.

We train the network in two steps: one local step, where each channel is trained separately with specific data. After that, the Cross-channel is trained with a complex dataset, containing audio with speech and music together. Our final architecture has two channels, each one with the same topology as the specific architectures. The Cross-channel has one convolution operation, without pooling, with 30 filters, each with a dimension of 2x2. The Cross-channel is connected to a hidden layer with 500 units.

To be able to use the Cross-channel, both channels must output data with the same dimensions. Our results showed that resizing the Music-specific channel output produced better performance. This can be explained by the fact that the Speech-specific channel depends strongly on the non-locality of the features. So, we resize the output of the music-specific channel before the Cross-channel to a dimension of 35x2. Figure 3 illustrates the final architecture.

III. Experiments

To evaluate our model on auditory emotion recognition, we use three different datasets and perform a series of experiments. We evaluate each of the three presented architectures with all the datasets. For each architecture we performed several experiments to find the best parameters. As we are talking about an elevated number of parameters to tune, our search was systematic and we did not do any parameter fine tuning. With that, our goal is to show how the proposed models behave in the presented tasks, leaving space for improvements on the choice of the parameters.

A. Datasets

The first dataset used was the GTZAN database [31]. This corpus is not directly related to emotion recognition, but to music genre classification. The task of music genre classification is similar to music emotion classification [32], because the idea is to cluster audio segments which are closely related based on auditory features. Music genres can also be used for emotion classification, since for example blues and soul music is more related to sadness or lonely feelings, and pop music more to happiness [32]. This database contains 1000 songs, each one with 30 seconds and a sampling rate of 22050 Hz at 16 bit, divided into ten musical genres: Blues, Classical, Country, Disco, Hip hop, Jazz, Metal, Pop, Reggae and Rock.

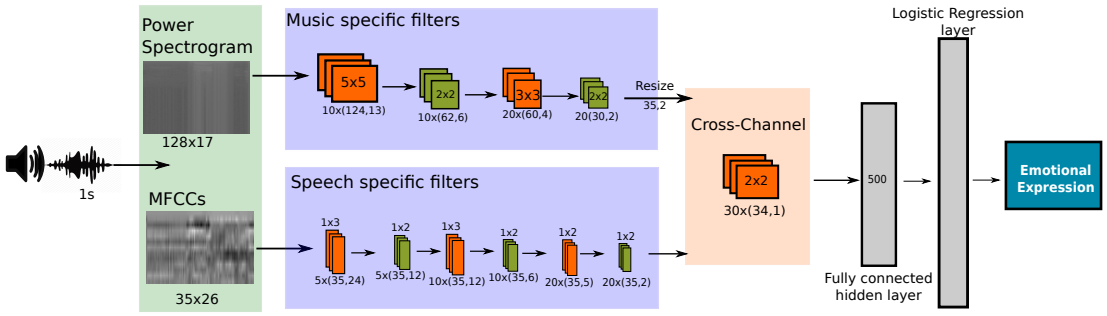

Fig. 3. Cross-channel architecture. In this architecture we have two channels, each one with the filters trained by the specific architectures. The output of each channel is conduced to a Cross-channel which is then trained. The Cross-channel learns how to correlate the two information and improve the generalization of the model. The Cross-channel has 30 filters, with dimension of 2x2.

The second database is the Surrey Audio-Visual Expressed Emotion (SAVEE) Database [33]. This corpus contains speech recordings from four male native English speakers. Each speaker reads sentences which are clustered into seven different classes: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise. Each speaker recorded 120 utterances, with 30 Neutral and 15 for each of the other emotions. All the texts are extracted from the TIMIT dataset and are phonetically-balanced.

The third database is the corpus for the Emotion-Recognition-In-the-Wild-Challenge (EmotiW). This corpus contains video clips extracted from random movies. This challenge is recognized as one of the most difficult tasks for emotion recognition, because the movie scenes can contain speech, music, sound effects, more than one speaker and even animals. We compare our results on this dataset only with the reported results using audio-only features, and not audio-visual information.

B. Methodology

For all datasets we extracted 1s of each audio input to train our networks. To recognize the audio, we used a sliding window approach of 1s. So, if the original audio input has 30s, we separated the audio in 30 parts of 1s and ran them through a network. With 30 results, we identified the most frequently occurring ones, leading to a final classification result for the 30s audio input.

For each dataset, we followed different protocols, established by the publisher of each corpus. This way we can compare our performance with published results for the same datasets. For the GTZAN dataset, we used a 10-fold cross validation approach. For the SAVEE dataset, we used a 4-fold cross validation strategy. The EmotiW dataset is used for the competition with the same name. The organizers made available a separated dataset for training and validation, but the testing set is available only for the ones which participate in the competition. As we did not have access to the testing data, we evaluated our model using only the validation set.

We performed experiments using the three architectures on all datasets. This allows us to explain the behavior of the model when non-suitable information is used as input for each task.

TABLE I
AVERAGE ACCURACY AND STANDARD DEVIATION FOR ALL THE EXPERIMENTS USING THE GTZAN DATASET.

Experiment	Accuracy (STD)	
	Trained	Pre-Trained
Music-Specific	96.4% (+/- 3.4)	-
Speech-Specific	68.7% (+/- 3.2)	62.5%(+/- 1.6)
Cross-channel	83.9% (+/- 2.3)	90.5%(+/- 2.2)

We also performed a Pre-training strategy, where the Music-specific architecture was trained exclusively with the GTZAN set, the Speech-specific architecture with the SAVEE set and the Cross-channel architecture uses the pre-trained features of both previous architectures and trains its own Cross-channel with the EmotiW corpus. This way we ensure that the Cross-channel architecture uses the specific representation learned by the specific architectures to construct a higher abstraction level of auditory features. The mean accuracy and standard deviation over 30 training runs are calculated for all the experiments.

C. Results

1) Music GTZAN dataset: Our Music-specific architecture obtained the best accuracy, with a total of 96.4%. The second best result appeared when using the pre-trained filters on the Cross-channel architecture, with a total of 90.5%, still almost 6% less than using only the Music-specific architecture. Using the Speech-specific architecture, the accuracy was the lowest, reaching the minimum score of 62.5% when applying the pre-training strategy. Table I exhibits all the experimental results on the GTZAN dataset.

2) Speech SAVEE dataset: On the SAVEE dataset, the Speech-specific architecture was the one which obtained the best mean accuracy, 92.0%. It was followed closely by the pre-trained version of the Cross-channel architecture, with 87.3%. The trained version of the Cross-channel obtained a total of 82.9%. Here the Music-specific architecture obtained the worst results, with a minimum of 63.1% on the trained version. The pre-trained version obtained slightly better results, reaching 64.5%. Table II exhibits all the experimental results on the SAVEE dataset.

TABLE II
AVERAGE ACCURACY AND STANDARD DEVIATION FOR ALL THE EXPERIMENTS USING THE SAVEE DATASET.

Experiment	Accuracy (STD)	
	Trained	Pre-Trained
Music-Specific	63.1% (+/- 2.7)	64.5% (+/- 2.3)
Speech-Specific	92.0% (+/- 3.9)	-
Cross-channel	82.9% (+/- 2.0)	87.3%(+/- 1.8)

TABLE III
AVERAGE ACCURACY AND STANDARD DEVIATION FOR ALL THE EXPERIMENTS USING THE EMOTIW DATASET.

Experiment	Accuracy (STD)	
	Trained	Pre-Trained
Music-Specific	22.1% (+/- 1.4)	23.1% (+/- 2.2)
Speech-Specific	21.7% (+/- 2.3)	21.0%(+/- 1.2)
Cross-channel	22.4% (+/- 1.1)	30.0%(+/- 3.3)

3) EmotiW dataset: The EmotiW dataset was to be the most challenging one. The pre-trained version of the Cross-channel architecture obtained the highest mean accuracy of 30.0%. All the other combinations, including the trained version of the Cross-channel, obtained accuracies around 20%. See table III for the results of our experiments with the EmotiW dataset.

IV. DISCUSSION

After performing all experiments, we were able to observe how each of our architectures behaved for each type of data. We could see that when we train and evaluate the architecture that is optimized for its corresponding corpus, the results were best. That means that using the Music-specific architecture produced best results for the GTZAN dataset and the Speech-specific architecture for the SAVEE dataset. It was possible to see that, except for the Cross-channel learning, the use of the pre-trained strategy for the specific architectures did not produce better results. This can be explained by the fact that when the pre-trained strategy was applied, the models actually learned very specific representations from their own dataset, and therefore lost their generalization capability. The application of the Cross-channel architecture with the pre-trained strategy actually damped that effect. In all datasets, the cross-channel architecture with the pre-trained strategy obtained results which were not the best, but close, with a reduction of less than 6% in the worst case (the GTZAN dataset). This shows that the pre-trained Cross-channel architecture lead to a generalization capability of the network, by recognizing data of different nature almost as good as a specific architecture. This behavior explains why the Cross-channel architecture with pre-training was able to obtain the best results on the EmotiW dataset.

The results obtained by our architectures are not far away from the state-of-the-art results in the literature. For the GTZAN dataset, our specific architecture performs close to the system proposed by Sarkar et al [34]. Their approach uses Empiric Mode Decomposition to compute pitch-based features from the audio input. They classify their features using a multilayer perceptron. Following the same evaluation protocol we used, they could reach a total of 97.70% of accuracy, slightly more than our 96.4% with the Music-specific architecture. Our Cross-channel architecture, when using the pre-training strategy, obtained a lower accuracy, but still competitive when compared to other results using different approaches. Table IV exhibits the results obtained on the GTZAN dataset. All the proposed techniques use a combination of several features, and a generic classifier such as SVM or MLP. However, using such a large number of audio features, their approaches are not suitable for generalization, a property that our Music-specific architecture has. The approach of Sgtia et al. is similar to ours. They evaluate the application of techniques such as dropout and Hessian Free training, but do not report the performance of the network neither for learning different features nor on generalization aspects.

TABLE IV
PERFORMANCE OF STATE-OF-THE-ART APPROACHES ON THE GTZAN DATASET. ALL THE EXPERIMENTS USE 10-FOLD CROSS VALIDATION AND CALCULATE THE MEAN ACCURACY. THE RESULTS OBTAINED BY SGTIA ET AL. [35] WERE USING A DIFFERENT DATA SPLIT, USING 50% OF THE DATA FOR TRAINING, 25% FOR VALIDATION AND 25% FOR TESTING.

Methodology	Accuracy(%)
Arabi et al. [36]	90.79
Panagakis et al.[37]	93.70
Sgtia et al.[35]*	83.0
Huang et al. [38]	97.20
Sarkar et al.[34]	97.70
Music-specific	96.40
Cross-Channel	90.50

For the SAVEE dataset, our approach is competitive. This dataset contains only speech signals, which are very different from the music signals. That explains the different accuracies obtained by the Music-specific architecture and the Speech-specific architecture. Once more the pre-trained Cross-channel architecture showed its generalization capabilities and was able to obtain a result which was comparable to the Music-specific architecture, having less than 3% of accuracy difference. When compared to state-of-the-art approaches, our Music-specific architecture obtained a result comparable with the work of Muthusamy et al. [39]. They use a particle swarm optimization technique to enhance the feature selection over a total of five different features, with many dimensions. They use an Extreme Learning Machine technique to recognize the selected features. Their work showed an interesting generalization degree, but still a huge effort is necessary, with the training step consuming enormous amounts of time and computational resources. The authors of the SAVEE dataset also did a study to examine the human performance for the same task. Using the same protocol, a 4-fold cross validation, they evaluated the performance of 10 subjects on the recognition of emotions on the audio data. The results showed that most approaches exceeded human performance on this dataset. Table V exhibits the state-of-art results and human performance on the SAVEE dataset.

The EmotiW dataset proved to be a very difficult challenge. On this dataset, our specific models did not work so well, but

TABLE V
PERFORMANCE OF STATE-OF-THE-ART APPROACHES ON THE SAVEE DATASET. ALL THE EXPERIMENTS USE 4-FOLD CROSS VALIDATION AND CALCULATE THE MEAN ACCURACY.

Methodology	Accuracy(%)
Banda et al. [40]	79.0
Fulmare et al.[41]	74.39
Haq et al.[42]	63.0
Muthusamy et al. [39]	94.01
Speech-specific	92.0
Cross-Channel	87.3
Human Performance [33]	66.5

TABLE VI
PERFORMANCE OF STATE-OF-THE-ART APPROACHES ON THE EMOTIW DATASET. ALL THE RESULTS CALCULATE THE MEAN ACCURACY ON THE VALIDATION SPLIT OF THE DATASET.

Methodology	Accuracy(%)
Liu et al. [43]	30.73
Kahou et al.[44]	29.3
Baseline results[45]	26.10
Cross-Channel	30.0

as Table VI shows this is also a much harder task. Due to the huge variability of the data, neither of them was able to learn strong and meaningful features by itself. When the Cross-channel architecture was used with the pre-training strategy, the network was able to learn to correlate the features of each channel, and use them to overcome the complexity of the dataset. Our Cross-channel architecture results are competitive with the state-of-the-art approaches, and performed better than the baseline values for the competition. The works of Liu et al. [43] and Kahou et al. [44] extract more than 100 auditory features each, and use classifiers such as SVM or multi-layer perceptrons to classify them. Our Cross-channel architecture results showed that we can actually obtain similar generalization capability using a simple and direct pre-training strategy without the necessity of relying on several different feature representations. Table VI exhibits the results on the EmotiW dataset.

Our experiments showed how the pre-training strategy improved the generalization of our Cross-channel architecture. It got comparable results of specific feature representation architectures, without relying on several different techniques to construct various features to be extracted. Our Cross-channel architecture also presented good results compared to state-of-the-art approaches. Our Cross-channel still is a deep neural network, and need and can handle a huge amount of data to show its best performance. The use of MFCCs and power spectrograms to represent specific features proved to be a valid idea, but still limits the information that the audio signal is carrying. A study over different feature representations, and maybe the combination of more features, could improve the generalization capabilities of our network even further.

V. CONCLUSION AND FUTURE WORK

Auditory emotion recognition is a very hard task, even for humans. Automatic systems are still not good enough to be used in natural scenarios, but demonstrated improvements in the last years. In this paper we showed three architectures for learning auditory information for emotion recognition tasks. Our architectures use specific information to generate general knowledge, and were shown to be competitive with different emotion recognition systems proposed in the last years.

Our two feature-specific architectures, one for music and one for human speech, performed very well on their tasks. However, when evaluated on their generalization capabilities, the performance was not optimal. The idea of the pre-trained Cross-channel architecture is to use the features learned by the specific architectures and learn how to combine them into a general emotion recognition system. The Cross-channel architecture was shown to be competitive and efficient in the three tasks we proposed: only music recognition, only speech recognition and a very complex scenario with audio clips extracted from movies.

Our networks were successful to some extent, but future work on learning what knowledge the networks contains can improve our comprehension of the architectures themselves and also of the auditory emotion recognition tasks. Also coupling these networks with visual deep networks could enhance their power for automatic emotion recognition. Finally, we expect another boost in performance by using recurrent neural networks for data with time-dependent constraints.

ACKNOWLEDGEMENTS

This work was partially supported by the CAPES Brazilian Federal Agency for the Support and Evaluation of Graduate Education (p.n.5951–13–5), the German Research Foundation DFG under project CML (TRR 169), and the Hamburg Landesforschungsförderungsprojekt.

REFERENCES

[1] F. Foroni and G. R. Semin, "Language that puts you in touch with your bodily feelings the multimodal responsiveness of affective expressions," *Psychological Science*, vol. 20, no. 8, pp. 974–980, 2009.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, Jan 2001.

[3] P. Rani and N. Sarkar, "Emotion-sensitive robots - a new paradigm for human-robot interaction," in *Humanoid Robots, 2004 4th IEEE/RAS International Conference on*, vol. 1, Nov 2004, pp. 149–167 Vol. 1.

[4] D. Bandyopadhyay, V. C. Pammi, and N. Srinivasan, "Chapter 3 - role of affect in decision making," in *Decision Making Neural and Behavioural Approaches*, ser. Progress in Brain Research, V. C. Pammi and N. Srinivasan, Eds. Elsevier, 2013, vol. 202, pp. 37 – 53.

[5] M. Cabanac, "What is emotion?" *Behavioural Processes*, vol. 60, no. 2, pp. 69 – 83, 2002.

[6] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," *Handbook of affective sciences*, pp. 433–456, 2003.

[7] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–17–IV–20.

[8] B. Schuller, J. Dorfner, and G. Rigoll, "Determination of nonprototypical valence and arousal in popular music: features and performances," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, p. 5, 2010.

[9] J. Madsen, B. S. Jensen, and J. Larsen, "Learning combinations of multiple feature representations for music emotion prediction," in *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, ser. ASM '15. New York, NY, USA: ACM, 2015, pp. 3–8.

[10] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller, "Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 473–480.

[11] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 4749–4753.

[12] M. Liu, H. Chen, Y. Li, and F. Zhang, "Emotional tone-based audio continuous emotion recognition," in *MultiMedia Modeling*, ser. Lecture Notes in Computer Science, X. He, S. Luo, D. Tao, C. Xu, J. Yang, and M. Hasan, Eds. Springer International Publishing, 2015, vol. 8936, pp. 470–480.

[13] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[14] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. rahman Mohamed, G. Dahl, and B. Ramabhadran", "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks "*, vol. 64", no. ", pp. 39 – 48", "2015", special Issue on Deep Learning of Representations".

[15] T. L. Li, A. B. Chan, and A. Chun, "Automatic musical pattern feature extraction using convolutional neural network," in *Proc. Int. Conf. Data Mining and Applications*, 2010.

[16] J. Schluter and S. Bock, "Improved musical onset detection with convolutional neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6979–6983.

[17] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[18] P. Barros, D. Jirak, C. Weber, and S. Wermter, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," *Neural Networks*, vol. 72, pp. 140 – 151, 2015, neurobiologically Inspired Robotics: Enhanced Autonomy through Neuromorphic Cognition.

[19] P. Barros, C. P Webber, and S. Wermter, "Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction," in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 646–651.

[20] B. Fasel, "Head-pose invariant facial expression recognition using convolutional neural networks," in *Proceedings of the fourth IEEE International Conference on Multimodal Interfaces, 2002*, 2002, pp. 529–534.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, p. 2012.

[22] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurons in the cat's striate cortex," *Journal of Physiology*, vol. 148, pp. 574–591, 1959.

[23] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, G. J. Gordon and D. B. Dunson, Eds., vol. 15. Journal of Machine Learning Research - Workshop and Conference Proceedings, 2011, pp. 315–323.

[24] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[25] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2007, pp. 1–8.

[26] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, Mar. 2010.

[27] J. George and L. Shamir, "Unsupervised analysis of similarities between musicians and musical genres using spectrograms," *Artificial Intelligence Research*, vol. 4, no. 2, p. p61, 2015.

[28] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

[29] D. Hau and K. Chen, "Exploring hierarchical speech representations with a deep convolutional neural network," *UKCI 2011 Accepted Papers*, p. 37, 2011.

[30] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 10, pp. 1533–1545, 2014.

[31] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE transactions on*, vol. 10, no. 5, pp. 293–302, 2002.

[32] S. M. Kim, A. Valitutti, and R. A. Calvo, "Evaluation of unsupervised emotion models to textual affect recognition," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, 2010, pp. 62–70.

[33] S. Haq and P. Jackson, *Machine Audition: Principles, Algorithms and Systems*. Hershey PA: IGI Global, Aug. 2010, ch. Multimodal Emotion Recognition, pp. 398–423.

[34] R. Sarkar and S. K. Saha, "Music genre classification using emd and pitch based feature," in *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*. IEEE, 2015, pp. 1–6.

[35] S. Sigtia and S. Dixon, "Improved music feature learning with deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6959–6963.

[36] A. F. Arabi and G. Lu, "Enhanced polyphonic music genre classification using high level features," in *Signal and Image Processing Applications (ICSIPA), 2009 IEEE International Conference on*. IEEE, 2009, pp. 101–106.

[37] Y. Panagakis and C. Kotropoulos, "Music genre classification via topology preserving non-negative tensor factorization and sparse representations," in *Acoustics speech and signal processing (ICASSP), 2010 IEEE international conference on*. IEEE, 2010, pp. 249–252.

[38] Y.-F. Huang, S.-M. Lin, H.-Y. Wu, and Y.-S. Li, "Music genre classification based on local feature selection using a self-adaptive harmony search algorithm," *Data & Knowledge Engineering*, vol. 92, pp. 60–76, 2014.

[39] H. Muthusamy, K. Polat, and S. Yaacob, "Particle swarm optimization based feature enhancement and feature selection for improved emotion recognition in speech and glottal signals," *PloS one*, vol. 10, no. 3, p. e0120344, 2015.

[40] N. Banda and P. Robinson, "Noise analysis in audio-visual emotion recognition," in *Proceedings of the 11th International Conference on Multimodal Interaction (ICMI)*. Citeseer, 2011.

[41] N. S. Fulmare, P. Chakrabarti, and D. Yadav, "Understanding and estimation of emotional expression using acoustic analysis of natural speech," *International Journal on Natural Language Computing (IJNLC)*, vol. 2, no. 4, 2013.

[42] S. Haq, P. J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition." in *AVSP*, 2009, pp. 53–58.

[43] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 494–501.

[44] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 543–550.

[45] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 461–466.