Neural Computing and Applications (2020) 32:1007–1022
https://doi.org/10.1007/s00521-018-3869-3

DEEP LEARNING FOR MUSIC AND AUDIO

Understanding auditory representations of emotional expressions with neural networks

Iris Wieser[1] · Pablo Barros[1] · Stefan Heinrich[1] · Stefan Wermter[1]

Received: 2 December 2017 / Accepted: 9 November 2018 / Published online: 11 December 2018

Abstract
In contrast to many established emotion recognition systems, convolutional neural networks do not rely on handcrafted features to categorize emotions. Although achieving state-of-the-art performances, it is still not fully understood what these networks learn and how the learned representations correlate with the emotional characteristics of speech. The aim of this work is to contribute to a deeper understanding of the acoustic and prosodic features that are relevant for the perception of emotional states. Firstly, an artificial deep neural network architecture is proposed that learns the auditory features directly from the raw and unprocessed speech signal. Secondly, we introduce two novel methods for the analysis of the implicitly learned representations based on data-driven and network-driven visualization techniques. Using these methods, we identify how the network categorizes an audio signal as a two-dimensional representation of emotions, namely valence and arousal. The proposed approach is a general method to enable a deeper analysis and understanding of the most relevant representations to perceive emotional expressions in speech.

Keywords Auditory emotion categorization · Affect analysis · Dimensional emotions · Deep neural network

1 Introduction

Perceiving and expressing emotions are key elements of natural communication. Constantly, humans relate to each other by conveying their emotional states and by reacting according to the perceived emotions [46]. Weninger et al. [64] claim that one of the most important factors in human communication is the emotional expression in sound. Using speech, humans transmit affective information implicitly via acoustic messages as well as explicitly via linguistic messages [42].

A speech emotion recognition system typically consists of a feature extracting method, which identifies the most relevant representations of the audio data, and a classifier, which identifies the emotion perceived in the corresponding speech utterance. There is evidence that a well-chosen selection of features improves the classification accuracy in emotion recognition tasks [16, 52].

Despite the fact that there exists a large amount of literature that relates handcrafted acoustic and prosodic features in speech to emotional states, e.g., [10, 52], it is not yet fully understood which features are most relevant to express emotional states [55]. The findings vary in emotion definitions, emotion categories, and the extracted acoustic features. Additionally, there exist some contradictory findings on the most influential acoustic features for emotional expressions [10]. This makes it difficult to select appropriate features for an emotion classification task.

Many emotion recognition systems currently extract a large amount of handcrafted acoustic and prosodic features that have originally been designed for speech recognition tasks [7, 47]. Although successful in constrained emotion recognition scenarios, such models are not scalable and fail when applied to real-world-related tasks [49]. Furthermore,

✉ Iris Wieser
4wieser@informatik.uni-hamburg.de

Pablo Barros
barros@informatik.uni-hamburg.de

Stefan Heinrich
heinrich@informatik.uni-hamburg.de

Stefan Wermter
wermter@informatik.uni-hamburg.de

[1] Knowledge Technology Group, Department of Informatics, Universität Hamburg, Hamburg, Germany

it has been shown that the most relevant features for acoustic emotion recognition depend on the used dataset, classifier, and emotional representations [36].

The problem, commonly found in many approaches in the area of emotion representations, is that researchers often consider emotions as discrete categories, such as anger or fear [10, 15]. However, this limits the number of possible emotional states that can be described. To allow a more flexible interpretation of emotional states, a dimensional representation of emotions can be used. A dimensional model offers a more fine-grained sentiment representation by describing emotional expressions with continuous values of various dimensions. The most common used dimensions are the two dimensions: *valence* and *arousal*. *Valence* is described by the pleasantness of a stimulus and represents the positive and negative feelings of a speaker, while *arousal* refers to the amount of energy used to express a specific emotion. Thus, arousal captures how reactive the subject is to a stimulus. Anger, for example, is represented by a rather high arousal and low valence. This two-dimensional model has become well-accepted [20], as it allows the distinction between describing the speech signal and associating it with different emotional states.

Besides emotion representation, another important challenge in the field of affective computing is the acoustic description of speech. In the past, several feature extractors have been proposed to represent auditory signals. Most of the work on affective computing is using various handcrafted descriptors [7, 18]. Despite the fact that handcrafted descriptors are commonly used, they might not represent the emotional characteristics in speech sufficiently and efficiently enough [16]. Instead of using many handcrafted features, generalization can be increased by learning auditory representations based on the data distribution. Deep neural networks (DNN) have been proposed in emotion-related tasks to automatically learn representations directly from audio signals for certain tasks [30, 35, 38, 58, 66]. One of the strategies used by the deep learning community is to bootstrap the learning process by using a spectral representation of the audio, usually with the use of spectrograms or Mel Frequency Cepstral Coefficients (MFCCs) [11]. Although more robust than the strictly handcrafted features [30, 67], such spectral representation loses important information regarding prosodic features [22].

One way to not limit how a deep neural network can learn emotional features is proposed by Trigeorgis et al. [62]. They suggest an end-to-end learning strategy, where the network learned to categorize emotions on unprocessed speech signals. In contrast to feature learning methods that are based on MFCCs or power spectrograms, the high resolution of the speech signals in time is preserved. The outstanding performance of the proposed network indicates that the detailed information of the raw and unprocessed audio signal in the time domain might allow the network to learn new and perhaps even complementary features compared to the handcrafted features which are used in most emotion recognition systems so far. Inspired by Trigeorgis et al., recent models also make use of unprocessed speech signals for emotion recognition [8, 23, 31]. They were able to achieve state-of-the-art performance on unconstrained speech emotion recognition tasks, moving one step closer to applications on real-world scenarios.

Although presenting impressive performance, most of the recent solutions lack an explanation on why their model works. By identifying how their models learn emotional features would give them the ability to adapt and improve their solution to be used on more challenging tasks. Trigeorgis et al. draw a direct comparison between the activations of their network's temporal gates and the standard handcrafted features such as loudness, energy, and pitch. A deeper analysis of their network would help to provide a better understanding of their success and extrapolate their findings to the context of their scenario and the used dataset. The analysis proposed by Trigeorgis et al. also would not be useful for other approaches, as their methodology is highly tailored to be used in their specific network. All recent solutions would benefit from an integrated analysis methodology that helps to understand and compare what emotional features have been learned by each of these solutions. As they can be complementary, such methodology could be the basis for a better understanding of this field and the proposed contributions.

In this paper, we aim to provide a general method that allows a deeper analysis of the learned acoustic features to improve the understanding of emotional expressions in speech. Our proposed architecture consists of a convolutional neural network (CNN) that is trained on raw auditory data to learn emotional characteristics. Our approach is an end-to-end learning method based on the approach of Trigeorgis et al. [62]. This means the network learns to classify emotions in the two-dimensional representation of valence and arousal directly from raw auditory information. To evaluate our model, we provide a comparison with state-of-the-art implementations for acoustic emotion recognition by using the interactive emotional dyadic motion capture (IEMOCAP) dataset. Moreover, we introduce two novel visualization methods that allow an interpretation of what our model has learned, i.e., how our model represents emotional characteristics from auditory information. Finally, we discuss how the network learns to identify arousal and valence on speech signals and how this is related to the fields of phonetics and descriptive linguistics.

2 Proposed method

The approach of this work is to implicitly learn acoustic and prosodic features directly from the raw audio signal and analyze the learned representations. The hypothesis is that this will increase the understanding of the features that influence the perception of emotional expressions in speech most.

In order to test this hypothesis, an architecture is proposed that consists of a convolutional neural network (CNN) to automatically learn the characteristic representations directly from the raw audio data. To analyze and understand the network's learned representations, two different methods have been applied: a performance comparison and several visualization techniques. The performance of the proposed network is compared to the following three approaches: First, state-of-the-art handcrafted features will be added to the architecture to investigate if this increases the performance. This might give some insight on whether the extracted features of the CNN on the raw audio signal complement other classical features. Second, feature learning is applied to the commonly used MFCCs only. This might show if the time resolution in the raw waveform carries important information that is discarded in the MFCC representation. Third, a multi-layer perceptron (MLP) is trained only on commonly used handcrafted acoustic features to evaluate the performance of the CNN in contrast to classical approaches.

Moreover, techniques have been adopted to analyze the learned representations of raw audio data. This analysis will provide insights on the most salient auditory representations of speech that are learned by the CNN to predict vocal emotional expressions.

2.1 End-to-end learning network

Representation learning, also referred to as feature learning, allows to automatically extract and organize discriminative information from data. The term *end-to-end* learning emphasizes that a system learns all parameters from raw and unprocessed data to predict the final output in one processing step [44]. This includes that the system does not require any explicit feature extraction methods, but automatically learns the internal representations required for the processing task. As an end-to-end learning method removes the need for the explicit design of classical features or selection methods, it also decreases the prior knowledge and engineering effort required to solve a problem.

Therefore, the proposed architecture uses end-to-end learning to automatically identify salient representations in audio signals for emotion-related tasks. By training the network on raw audio data, it might even learn new and unexplored auditory features. In the following, the proposed end-to-end learning architecture will be referred to as *ELoR network* (end-to-end learning on raw audio).

2.1.1 Architecture

The ELoR network should provide a feature extraction method that is capable of extracting higher-level representations of the input signal and a classifier that predicts the emotional state expressed in the input signal.

For the feature extraction, supervised training of a CNN [39] will be used to provide a generalized model that is capable of learning relevant acoustic features. Since their introduction by LeCun et al. [40] in 1989, CNNs have shown outstanding performance on various tasks, such as handwritten digit classification [41] and image classification [9, 37]. CNNs are feedforward artificial neural networks that are inspired by the locally sensitive orientation-selective cells of the primary visual cortex [32]. A CNN consists of two alternating types of layers: convolutional layers and pooling layers. Figure 1 illustrates the basic principles of the convolutional and pooling layers of a CNN.

The *convolutional* layers use convolution operations on local receptive fields by using filters to extract higher-level features. In Fig. 1, three filters with the size of 3×3 are used for the first convolutional layer. The values of the filters are convolved with the original matrix, which means they are multiplied element-wise on the original matrix and summed up. To get a full feature map of a convolutional layer, the filter corresponding to that feature map (illustrated by the colors red, green, and orange, respectively) is slid over the complete input matrix. Then, for the second convolutional layer, the input volume has increased from one input matrix to three input matrices that correspond to the number of maps obtained in the previous convolutional layer. To generate each of the five feature maps, the sum of convolutions for each filter per input feature map is taken (illustrated, for example, in Fig. 1 in blue).

The activation $a_{k,i}$ of one unit, also referred to as neuron, at (x, y) of the i-th feature map within the k-th layer can be described by

$$a_{k,i}(x,y) = f\left(b_{k,i} + \sum_{j=1}^{N_{(k-1)}} (w_{k,i,j} * a_{(k-1),j})(x,y)\right), \qquad (1)$$

where $b_{k,i}$ represents the shared bias for the i-th feature map of the k-th layer. $N_{(k-1)}$ is the number of input matrices in the layer $k-1$. The shared weight matrix $w_{k,i,j}$ of the i-th feature map in the k-th layer with a size of $H \times L$ is to be convolved with the input matrix j. The activation

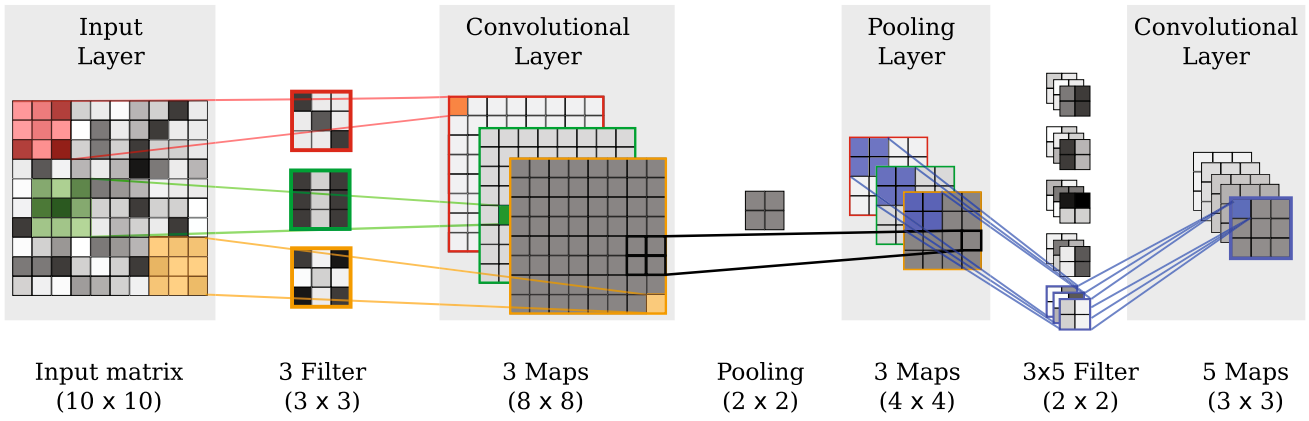

Fig. 1 Basic principles of a CNN showing a two-dimensional input matrix, a convolutional layer with 3 filters of the size 3×3, a pooling layer with the size 2×2 and a convolutional layer with 5 filters of size 2×2. The computation of one exemplary unit of each feature map is indicated by the colors red, green, orange and blue. Each color corresponds to one feature map (color figure online)

function f used with CNNs is typically a rectified linear unit (ReLU) function, defined by $f(x) = \max(0, x)$ [25]. It has been shown that the ReLU function increases the convergence of stochastic gradient descent compared to other commonly used activation functions, such as the sigmoid function [37], due to their non-saturating characteristic.

The *pooling* layer uses a downscaling operation to reduce the spatial size of the feature maps generated by the convolution layer. By downscaling the feature maps, they also become less sensitive to specific locations of structures within the input signals. The most common downscaling function is max-pooling. For max-pooling, only the maximum activation of the local receptive field is passed to the next layer. In Fig. 1 the used pooling size is 2×2.

By learning only the shared weights (filters) per layer, CNNs greatly reduce the parameters that need to be learned. Moreover, CNNs are able to learn representations from spatial invariant low-level features. Due to their four main concepts of local receptive fields, shared weights, pooling operations and usage of many layers, CNNs are capable of learning appropriate feature extractors based on "raw" inputs in a supervised manner with backpropagation [39]. This will enable generalization capabilities and allow to automatically learn the best feature set in the context of emotional expressions.

The basic concept of the implemented ELoR network is shown in Fig. 2. The CNN part of the ELoR network consists of 6 layers in total. For the three convolutional layers, one-dimensional filters are applied to the discrete-time waveform $x(t)$ with a stride of 1. For all three pooling layers, max-pooling is applied to the feature maps generated by the corresponding convolutional layers. It has been shown that max-pooling performs best for feature extraction with CNNs on raw audio data [51]. After each combination of convolution and pooling layer, the nonlinear activation function ReLU is used, which is suggested to be analogous to a process in the human ear [62].

For weight initialization in these layers, a normalized initialization, also referred to as Xavier's initialization [24], is applied. As suggested by He et al. [28], a minor adaptation of the normalized initialization is additionally made to address the special characteristic of the ReLU activation function, which is zero for half of its input. Additionally, batch normalization [33] is used for regularization of the network. By normalizing the output of a convolutional layer, the number of required training steps can be decreased [33].

For classification, fully connected dense layers or multilayer-perceptron (MLP) layers will be used. Nummenmaa et al. [46] found some indications that different areas of the human brain are activated when perceiving valence and arousal in language. It seems that arousal is more correlated with the auditory cortices, Broca's area, thalamus, and amygdala, while valence is more correlated with the lateral frontal and orbitofrontal cortices. Therefore, each emotional dimension is classified separately.

As shown in Fig. 2, for each emotional dimension an individual MLP is used. Two MLPs are applied to the output of the CNN to predict the emotional state corresponding to the raw auditory input. Each MLP consists of two hidden layers. To prevent the network from overfitting, dropout [29] is utilized on each layer during training. The activation function for all layers except the output layer is a sigmoid function. For the layers in the MLPs, the normalized weight initialization is used, as suggested by Glorot and Bengio [24].

A softmax activation function is applied to each of the two output layers independently to obtain the output of the network. Given the input signal x, the softmax function

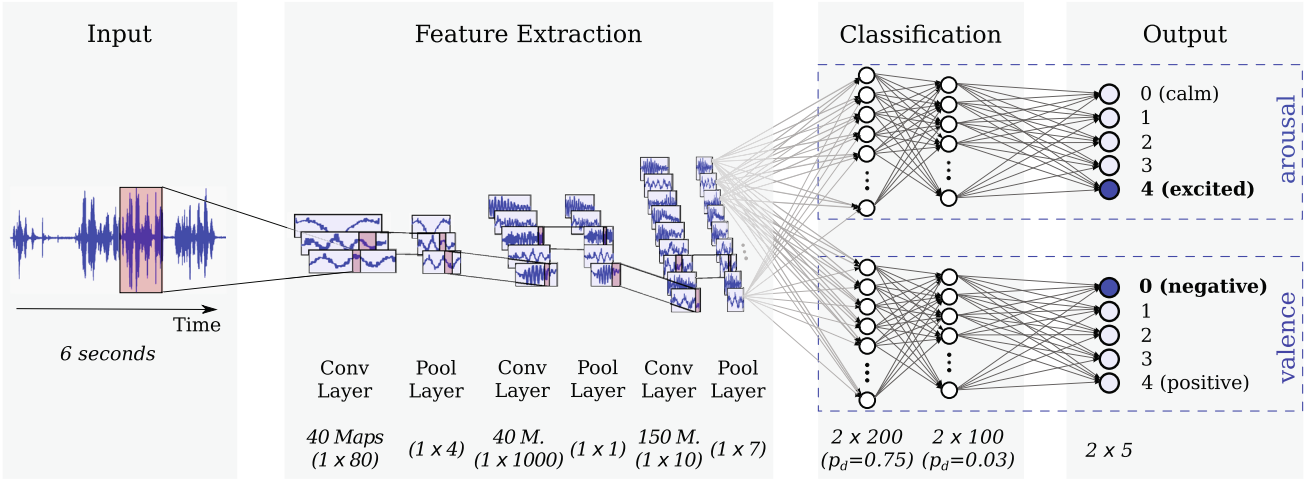

Fig. 2 End-to-end Learning on Raw auditory architecture. The ELoR network includes the raw auditory input of 6 seconds length, a CNN for feature extraction with 1D filters, two split MLPs for classification and 2×5 output neurons representing the annotations used in the dataset. The parameters found by a hyperparameter optimization are given in italic (see also Sect. 3.3)

calculates the probability distribution of the 5 output classes y_i per emotional dimension by

$$\text{softmax}(y_i|x) = \frac{e^{y_i(x)}}{\sum_{j=1}^{5} e^{y_j(x)}} \quad \text{for} \quad i = 1\ldots5. \tag{2}$$

The final prediction of the network is the two output labels with the maximum probability for each dimension.

The categorical cross-entropy loss has been employed to compute the gradients during backpropagation. It has been shown that by using the cross-entropy loss in multi-class classification problems, a local optimum can be reached more efficiently compared to the mean squared loss [26]. Additionally, for all layers an L2 weight constraint $||w||_2$ was implemented to keep the learned weights small, as suggested by Hinton et al. [29]. Thus, the final loss function used in this study is defined by

$$E(w) = -\sum_{j=1}^{N} y(x_j)\log(\hat{y}(x_j)) + \alpha_w||w||_2, \tag{3}$$

where the hyperparameter α_w defines how much the L2 loss influences the complete loss.

Stochastic gradient descent (SGD) with Nesterov momentum [45] is used during training, which has been shown successful for training deep neural networks [3, 61]. To train the complete network, the two MLPs are updated successively on the same input. First, the loss function for one label (e.g., arousal) is computed and the weights of the corresponding MLP and the CNN are adjusted to minimize the loss. Then, on the updated network the loss for the second label (now valence) is computed for the same input to adjust the weights of the second MLP and again the CNN. To avoid any influence of the order of weight updates, it is randomly chosen which MLP is updated first.

2.1.2 Input

The input of the ELoR network is raw audio data. By learning representations directly in the time domain, the high time resolution of the audio signal is preserved and might allow extracting more information over time compared to spectrograms. Each sample is represented by a one-dimensional vector that contains the amplitude of the signal over time. The samples given to the network need to be of equal length because the architecture proposed in the following has a fixed input size.

The best window length still seems to be an open question [67]. It has been shown that an emotion lasts around 500 ms to 4 s [14]. Moreover, it was claimed that most researchers focusing on vocal emotional expressions use windows of speech of approximately 2–6 s [27]. Pollack et al. [48] showed that it is even possible to classify 16 different modes of expressions (e.g., boredom, fear, uncertainty, and so on) within only 60 ms with a recognition accuracy of $\sim$50%.

Those findings led to the conclusion that an emotional state is sufficiently represented within a 6 s window of speech. As longer sequences would exceed the memory resources during the training phase in the experiments conducted within this study, each sample will be of 6 s length. Assuming a sampling rate of 16 kHz the size of an input array for one sample would be $1 \times 96{,}000$.

2.1.3 Output

The output of the network is emotional states. They are represented by the two dimensions valence and arousal. The annotations in the dataset used for the experiments are based on five symbols for each dimension. As each

dimension is split into discrete categories, the problem can be considered as a classification problem instead of a regression problem. To represent the concepts that have been used for annotations, the output is modeled by five output neurons for each dimension, as illustrated in Fig. 2. The modeled task can be described as a five-class two-label classification problem. In total, there exist 25 possible combinations of valence–arousal annotations.

2.2 Architectures for comparison

In order to compare the performance of the ELoR network, three additional architectures have been implemented. These architectures share the concept of the ELoR network but differ in the input representations and the feature extraction methods. This way a relative performance comparison can show whether the learned representations of the proposed ELoR network are competitive with state-of-the-art approaches. Moreover, a detailed analysis of the predictions will give some indications on whether the ELoR network is capable of learning new and unexplored features.

The additional architectures are called the *LoM network* (Learning on MFCCs), which learns the features only based on MFCC representations, the *FTR network* (FeaTuRes), which uses only handcrafted features to classify an emotional state, and the *ELoR + FTR network*, which extends the input of the ELoR network by additional handcrafted features. In the following, only the differences between each network and the ELoR network are explained in more detail.

2.2.1 LoM network—feature learning only on MFCCs

The input representations used for the LoM model are the first 26 MFCCs. They are computed on the 6-s audio samples with a sampling rate of 16 kHz. By using the standard window size of 25 ms, the commonly used shift of 10 ms and a frequency resolution of 1024 Hz, the dimension of the computed MFCC spectrogram is 599×26. The network needs to learn the representations of each MFCC independently to avoid learning correlations among the coefficients that might not exist [1]. Thus, each of the 26 coefficients is treated as an individual channel in the input layer (comparable with red, green, blue in visual tasks) and only one-dimensional filters are used for both convolution and pooling. Figure 3a illustrates the basic concept of the LoM network, which is proposed for feature learning based on MFCC representations only.

The relative performance comparison of the ELoR network and the LoM network is supposed to provide some indications whether additional or maybe even complementary representations can be extracted from the audio signal in the time domain compared to the frequency domain.

2.2.2 FTR network—handcrafted features only

For the FTR network, classic handcrafted features are manually extracted instead of automatically learned. The features have been computed from 6 s of a speech signals with the openSMILE feature extractor [18, 19]. For better reproducibility and comparability one of the standard acoustic feature sets presented by Eyben et al. [18] is chosen. The standard acoustic feature set for the INTERSPEECH 2010 Paralinguistics Challenge (IS10), originally defined by Schuller et al. [56], is a commonly used benchmark set for vocal emotional expression tasks [12, 13, 60]. For example, Jin et al. [34] also used this feature set for emotion recognition on the IEMOCAP dataset. Based on this acoustic feature set 1582 features are extracted. For each 6-s input sample, the 1582 features are extracted and directly fed to the two MLPs for classification. Figure 3b illustrates the basic concept of the FTR network, which is proposed to predict an emotion based on handcrafted features only.

The relative performance comparison of the ELoR network and the FTR network can show if the ELoR network is capable of learning representations that are competitive with state-of-the-art handcrafted features.

2.2.3 ELoR + FTR network

The ELoR + FTR network is a combination of the originally proposed ELoR network and the FTR network. For this network, the learned representations of the ELoR network are extended by the 1582 manually extracted features, proposed for the FTR network. Figure 3c illustrates the basic concept of the ELoR + FTR network, allowing a combination of end-to-end feature learning and manually extracted handcrafted features. Both, the learned representations of the CNN and the handcrafted features are fully connected to the two MLPs.

The relative performance comparison of the ELoR network and the ELoR + FTR network is suggested to indicate whether the implicitly learned features are complementary to state-of-the-art handcrafted features. If the combination of the handcrafted features and the learned representations improves the classification performance, the learned representations of the ELoR network might complement the handcrafted features.

2.3 Techniques to analyze learned representations

To visualize and understand the learned representations of the trained ELoR network on a higher level of abstraction, two different visualization methods are applied. This will provide some insights on the most salient auditory features

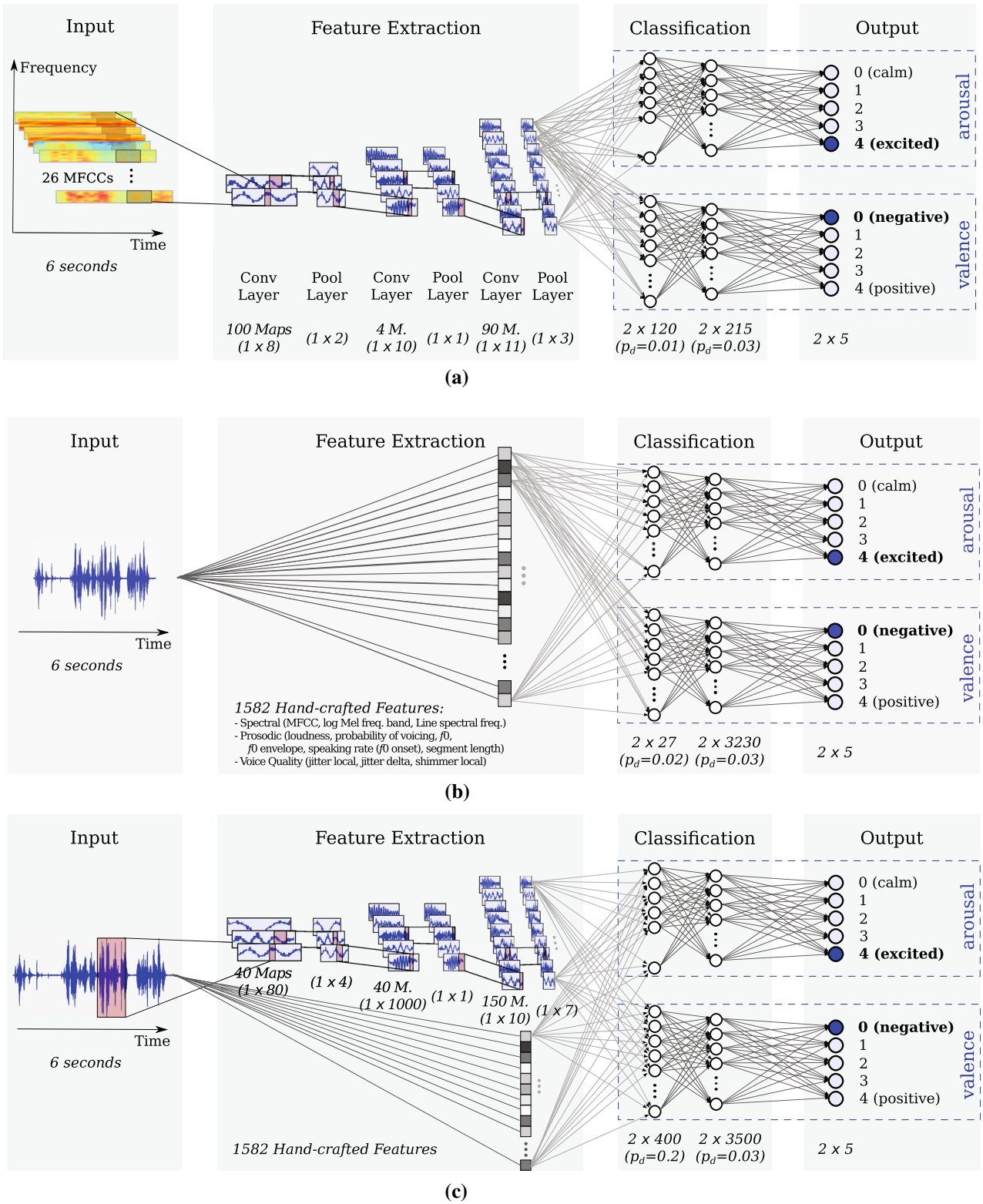

Fig. 3 Architectures for comparison, differing in the input representations and the feature extraction methods. Parameters found in hyperparameter optimization are given in italic (compare Sect. 3.3). **a** Feature learning only on MFCCs. The LoM network includes the MFCCs extracted on 6-s samples, a CNN for feature extraction with 1D filters, split MLPs for classification, and 2 × 5 output neurons. **b** Handcrafted features only. The FTR network includes 1582 manually extracted features on 6-s samples, split MLPs for classification, and 2 × 5 output neurons. **c** End-to-end learning with added features. The ELoR + FTR network includes the raw audio input of 6-s length, a CNN for feature extraction with 1D filters, 1582 manually extracted features as additional input for the split MLPs for classification, and 2 × 5 output neurons

that have been implicitly learned for vocal emotional expressions.

2.3.1 Network-driven visualizations

To achieve an understanding and a visualization of all learned layers in the CNN, the idea is to obtain a generalized input signal that maximizes the network's prediction for a specific class. This approach has been originally introduced by Erhan et al. [17]. It can be formulated as an optimization problem, where a randomly initialized input signal x_0 is updated by maximizing the activations for a specific class (output neuron) of a trained network with fixed parameters. Erhan et al. [17] applied this method on deep belief networks (DBNs) and stacked denoising autoencoders (SDAEs) to acquire a first-order representation of a neuron's behavior. Simonyan et al. [57] used this visualization technique also on a CNN for ImageNet classification. However, instead of using a randomly initialized input, they utilized a mean image resulting from the training set. Moreover, they added an L2 regularization term to prevent extreme single-pixel values and, thus, achieve more naturalistic images.

Given a trained network with fixed weights and biases and a class of interest, the input can be numerically calculated by a stochastic gradient ascent method. Let $a_{c,k}$ be the activation of the c-th neuron in the k-th layer, which represents the class of interest in the output, the generalized input x^* can be calculated by

$$x^* = \text{argmax}_x \left(a_{c,k}(x) - \lambda ||x||_2^2 \right). \tag{4}$$

This allows a visualization that is independent of a certain input but also represents the trained network completely.

2.3.2 Data-driven visualizations

For data-driven visualization techniques, the gradient of the network is computed with respect to a specific input signal. The obtained so-called saliency map highlights how much each given input value influences a specific neuron's activation. By calculating the saliency map of the predicted class for a certain input, it is possible to visualize the parts of the input that have the most impact on the network's final prediction. For the computation of the saliency maps, we used guided backpropagation [59], as it allows to obtain sharper saliency maps.

3 Experiments

For the evaluation of our architecture, the ELoR network is trained on the IEMOCAP dataset. Additionally, the three previously presented variants (the LOM network, the FTR network and the ELoR + FTR network) are trained on the same dataset to allow a comparison of their classification accuracy. The two visualization strategies presented in the previous section (network-driven and data-driven) are applied to the trained ELoR network to gain insights on which features are learned by the ELoR network to represent emotions.

A hyperparameter optimization was performed for each network to guarantee the competitiveness of each of the evaluated models. This ensures that our comparative evaluation is fair, as for each model the optimal parameters have been evaluated on the IEMOCAP dataset and were chosen for the final experiments. All the experiments were executed with the same hardware and are based on the same software libraries.

3.1 Dataset

For the training and performance analysis, the interactive emotional dyadic motion capture (IEMOCAP) dataset [6] is used. It is a multimodal and multi-subject acted dataset with improvisations and scripted dialogue scenarios to express various emotional states. It consists of approximately 12-h audiovisual data performed by ten actors. The actors have been separated for the dyadic conversations into pairs, consisting always of one male and one female. The data include speech and text transcriptions and are annotated with both categorical and dimensional labels.

Each recorded dialogue has been manually segmented using the dialogue turns to create continuous utterances, so that only one of the actors is speaking most of the time. The utterances vary in length from less than 1 s to 34 s. Each utterance has been annotated by two to three annotators on the dimensional scale on both auditory and visual information. Moreover, the clips have been shown to the annotators sequentially and thus with contextual information. For dimensional annotations, each dimension (valence, arousal, and dominance) is represented by a five-point scale.

On average, each sample in IEMOCAP has been annotated by 2.12 evaluators. Even though evaluators rarely annotated all sample, there is still a relatively high disagreement. For only 16% of all samples, the evaluators agreed on the exact same value combination for valence and arousal (with 25 combination possibilities in total). One reason for that might be that every human has a personal bias that depends on his culture, language, and personality. This is in agreement with a study, conducted by Metallinou and Naranyanan [43], that shows that humans perform better in rating emotions in relative terms instead of an absolute scale. Additionally, in some samples, we have found a disagreement of the evaluators as they have

set a different focus (e.g., one is rating the subjective feeling, while the other focuses on the mood or interpersonal stances). To minimize subjective confusion and bias, in this study only the evaluator "*E*-2" is considered for the evaluation. Although this evaluator annotated fewer samples than the others, the annotations are more balanced over the complete valence–arousal space and might allow the networks to learn a higher variety of emotional expressions.

3.2 Experimental setup

The input of the proposed model needs to have a fixed length for the CNN. Therefore, every sample has been sliced or padded with zeroes to a length of 6 s. For longer recordings than 6 s, every 6 s of the utterance are sliced and used as an individual sample with the same annotation. To avoid training on the last phoneme of an utterance exclusively, the last slice of a long utterance is used only if it is longer than 3 s. This results in 8522 samples for evaluator *E*-2.

Each sample is represented by a signed 16-bit integer with the values in the range of $[-32{,}768, 32{,}767]$. The amplitude of the input has been scaled down to 32-bit floating point values within the range of $[-1.0, 1.0]$. No further preprocessing steps have been applied to the recordings provided by the IEMOCAP dataset.

To allow an objective comparison of the three proposed networks, experiments are performed to evaluate the classification performance of the networks on the IEMOCAP dataset. All experiments have been conducted in a tenfold cross-validation evaluation scheme, where each fold consists of the samples of exactly one speaker. Thus, for each evaluation, the samples of nine speakers were distributed over the training and validation sets, while the remaining samples of the last speaker were used as test set. This allows a speaker-independent evaluation of the networks. For each network, the same 10 test sets have been used for evaluation.

Our visualization techniques have been applied to the trained ELoR network for the following representative test set: *Ses03-F*. This test set includes all samples of the female subject in the third session from the IEMOCAP dataset.

3.3 Hyperparamter optimization

The hyperparameters have been optimized for each architecture individually with the Python library hyperopt [5] to find the best possible configuration. For the hyperparameter optimization, the Tree of Parzen Estimator (TPE) [4] is used. It is an optimization strategy to find the hyperparameters which achieve the highest performance accuracy on a certain validation set within a defined search space. The validation set, on which the architectures have been optimized, consists of 20% randomly chosen samples of all samples considered from the IEMOCAP dataset in the respective cross-validation step.

For the ELoR network, the search space has been defined by a uniform distribution with a mean based on the hyperparameters suggested by Trigeorgis et al. [62]. In total, 70 evaluation phases have been performed to optimize the hyperparameters of the ELoR network. On an 8 GB GPU GeForce GTX 1080, the hyperparameter optimization for the ELoR network took 49 computing days. Table 1 and Fig. 2 show the results of the hyperparameter optimization and thus the final hyperparameters used for the ELoR architecture in the following experiments.

For the LoM network, the search space has been chosen based on the hyperparameters suggested by Barros et al. [2]. In total, 150 evaluation phases have been executed for the hyperparameter optimization on the LoM network. Table 2 and Fig. 3a illustrate the final hyperparameters selected for the conducted experiments with the LoM network.

For the FTR network also 150 evaluation phases have been executed for the hyperparameter optimization. Table 3 and Fig. 3b illustrate the final hyperparameters selected for the conducted experiments with the LoM network.

It was not feasible to find the optimal hyperparameters for the ELoR + FTR network, as the hyperopt search for the ELoR network took already 49 computing days. Therefore, combinations of the results for the ELoR

Table 3 Selected hyperparameters for the FTR network

Parameter	Value	Parameter	Value
Batch size	30	Learning rate	0.00004
Epochs	16	Learning decay	0.99
Momentum	0.65	L2 regular. α_w	3.4×10^{-8}

Table 1 Selected hyperparameters for the ELoR network

Parameter	Value	Parameter	Value
Batch size	15	Learning rate	0.006
Epochs	9	Learning decay	0.9
Momentum	0.65	L2 regul. α_w	1×10^{-9}

Table 2 Selected hyperparameters for the LoM network

Parameter	Value	Parameter	Value
Batch size	5	Learning rate	0.08
Epochs	14	Learning decay	0.75
Momentum	0.7	L2 regul.α_w	0.001

Table 4 Selected hyperparameters for the ELoR + FTR network

Parameter	Value	Parameter	Value
Batch size	15	Learning rate	0.005
Epochs	6	Learning decay	0.9
Momentum	0.65	L2 regul. α_w	1×10^{-9}

network and the FTR network have been empirically combined. Table 4 and Fig. 3c show the final hyperparameters selected for the ELoR + FTR network.

4 Results

First, the performance of the proposed architectures is presented and compared with different metrics to evaluate if the ELoR network is competitive with state-of-the-art approaches. Then, the results of different visualization techniques of the ELoR network are shown and explained in more detail to gain a better understanding of the learned representations.

4.1 Performance comparison

For each network, the predictions and targets of all 10 test sets have been combined to compute the average of the weighted average recall (WAR) and the unweighted average recall (UAR) [18] in the most unbiased manner over all test sets [21]. Table 5 presents the WAR and UAR for each network and each emotional dimension.

The performance results show that the proposed ELoR network is competitive with state-of-the-art approaches. For arousal, the ELoR network seems to perform best, whereas the FTR network seems to perform worst. This illustrates that the CNN might be able to extract information from the raw audio signal, which is not represented in the handcrafted features used for the FTR network.

Moreover, the results indicate that the ELoR + FTR network performs best on valence. This shows that for valence the learned representations of the ELoR might be complementary to the handcrafted features used for the FTR network.

As the IEMOCAP dataset is highly imbalanced, a classifier could achieve a higher WAR score than 0.200 (random guessing) by always predicting the most frequent label. Based on the prior probabilities of the test sets, the maximal WAR score for predicting only one label could be reached by always predicting arousal = 3. For valence, the maximal WAR score could be reached by only predicting valence = 1.

Table 5 shows that the scores for the weighted average recall (WAR) are on average higher than for the unweighted average recall (UAR). This indicates that all networks learned to consider the imbalances of the test sets and used this information to increase the prediction accuracy of an emotional state over all samples. Table 5 additionally illustrates that the classification accuracy on arousal is generally higher than on valence. This is in agreement with several studies showing that in speech usually the prediction accuracy for arousal is much higher than for valence [50, 54, 63, 65].

Table 5 Results on the average performance of each network

Model	Arousal		Valence	
	WAR	UAR	WAR	UAR
ELoR net	**0.524**	**0.394**	**0.429**	0.300
LoM net	0.510	0.351	0.424	0.281
FTR net	0.466	0.307	0.412	0.263
ELoR + FTR net	0.472	0.362	0.411	**0.302**
Random guessing	0.200	0.200	0.200	0.200
Most frequent label	0.408	0.200	0.392	0.200

The weighted (WAR) and unweighted (UAR) average recall are shown for each network. Additionally, the average results for random guessing classifiers are compared

Best values are given in bold

As the proposed ELoR network is competitive to networks with state-of-the-art handcrafted features, a deeper analysis of the learned representations of this architecture can give valuable information on the most important acoustic and prosodic features in vocal emotional expressions.

4.2 Learned representations

For a deeper understanding of the implicitly learned representations, two different visualization methods have been applied to the trained ELoR network: data-driven visualizations (saliency maps) and network-driven visualizations (generalized input signals).

4.2.1 Data-driven visualizations

The data-driven visualization method allows illustrating a saliency map based on the network's prediction on a specific speech input. The obtained saliency maps indicate which segments of an utterance have the most influence on the final prediction of the network. Figure 4 shows an exemplary saliency map that has been computed for the prediction of arousal based on a specific recording.

The Pearson correlation coefficient is used to identify how much the amplitude of the input influences the final prediction of the ELoR network. Based on the IEMOCAP

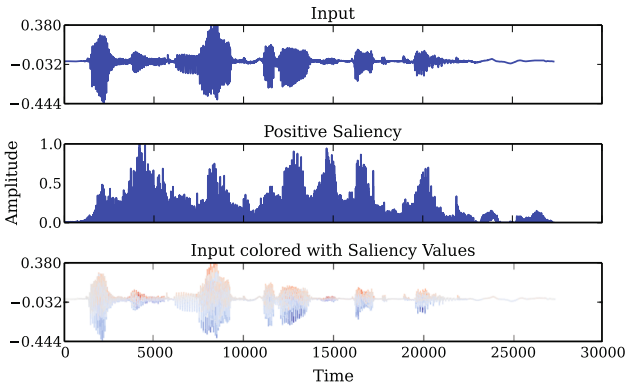

Fig. 4 Visualization of an exemplary saliency map that has been obtained for the prediction on arousal on a recording with a sampling rate of 16 kHz. The first row shows the input signal in the time domain. The second row illustrates the positive values of the obtained saliency map. The last row gives a visualization of the input signal, while the colors represent the values of the saliency maps at each time step. Red represents a high positive saliency value, while blue represents a high negative saliency value (color figure online)

dataset, the average correlation between the input signal and the corresponding saliency map for arousal is 0.50 and for valence 0.53. This shows that there exists a correlation between the amplitude of the speech signal and the obtained saliency map on both dimensions. This means that the amplitude of the signal in the time domain is important to predict the perceived emotional state of the speaker. Thus, it can be concluded that the louder a phoneme is perceived relative to the complete signal, the more influence it has on the final prediction of the emotional state. Since spectrograms do not explicitly represent the amplitude of the signal in the time domain, this also suggests that the ELoR network can extract relevant information, which is not represented in the MFCC representation used for the LoM network.

Another finding is that there exists a strong correlation between the saliency maps obtained for the prediction on valence and the saliency maps computed for the prediction on arousal for each sample of the IEMOCAP dataset (on average the Pearson correlation coefficient is 0.96). The saliency maps represent the gradients of the complete ELoR network. This includes the shared CNN network as well as the separated MLPs for each emotional dimension. As both saliency maps for valence and arousal have been obtained on the same input sample, the strong correlation suggests that the final prediction on both dimensions relies on the same parts of the input signal. Moreover, the strong correlation indicates that both MLPs seem to rely mainly on the same parts of the learned representations of the CNN. Therefore, it can be concluded that the perception of a specific emotional state depends on both dimensions on similar representations of the utterance.

4.2.2 Network-driven visualizations

The network-driven visualizations allow for a better understanding of the implicitly learned representations independent of specific input samples. Based on the fixed and already trained parameters of the ELoR network, input signals are learned to maximize a specific output (e.g., arousal = 0). For each label on both valence and arousal, a generalized input signal is obtained. As there exist five labels for each of the two dimensions arousal and valence, 10 generalized input signals have been learned in total.

The generalized signals for arousal and their characteristics are presented in Fig. 5a and Table 6 (left).

The visualizations of the learned input signals show that the ELoR network predicts more likely a higher degree of arousal, if

- the signal contains high frequencies;
- the signal contains many interruptions;
- the signal contains higher peaks of the amplitude; and
- the signal contains periodic frequencies for a longer period of time.

The results indicate that the perception of a more excited emotional state is correlated with higher frequencies, a faster-speaking rate, and longer utterances. The overall loudness of an utterance seems to have no influence on the prediction of arousal. This might be due to the fact that the IEMOCAP contains highly different speakers. For example, the female subject in the first session is very loud and extroverted, while the female subject in the third session seems to be rather shy and relatively quiet. However, the results suggest that higher peaks of the amplitude correlate with the perceived excitement. Thus, it seems that the maximal difference of amplitude is more important than the overall loudness for the prediction on arousal.

For valence, the generalized signals and their characteristics are presented in Fig. 5b and Table 6 (right).

The visualizations of the generalized input signals indicate that the ELoR network predicts more likely

- an extreme value (i.e., valence = 0 or valence = 4) if the signal contains high frequencies;
- an extreme value if the signal contains periodic frequencies for a longer period of time;
- a lower value (i.e., valence = 0 or valence = 1) if the signal exhibits clear and constant pauses in frequency and duration;
- a lower value if the signal is louder; and
- a higher value (i.e., valence = 3 or valence = 4) if the signal exhibits pauses that decrease in frequency and increase in duration.

For valence, the results indicate that the emotional state is perceived more negatively when the utterance exhibits

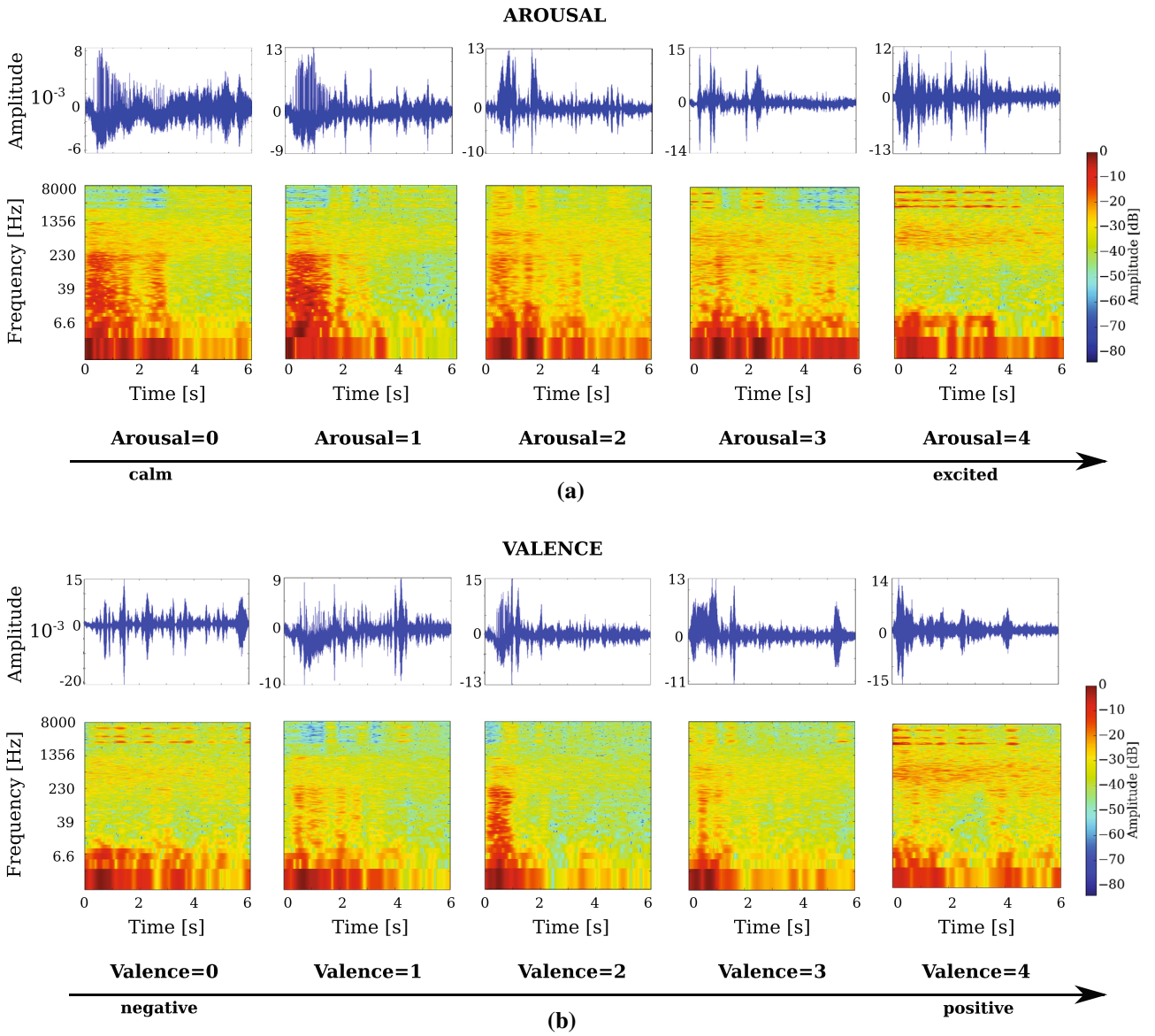

Fig. 5 Characteristics of the learned and generalized input signals that maximize a specific output in the ELoR network for each emotional dimension: arousal and valence. Each signal has been learned to maximize the network's prediction for that specific label. The signals are illustrated in both time domain (top) and frequency domain (below). **a** Generalized input signals for arousal in ascending order by the corresponding labels. **b** Generalized input signals for valence in ascending order by the corresponding labels (color figure online)

Table 6 Simplified and summarized results for all generalized input signals on arousal (left) and valence (right)

	Arousal					Valence				
	0	1	2	3	4	0	1	2	3	4
Loudness (dBFS)	− 57.3	− 55.6	− 56.7	− 58.3	− 56.5	− 55.7	− 56.6	− 56.8	− 57.0	− 57.6
Dom. f. (kHz)[a]	0–0.23	0–0.23	0–8.0	2.8–8.0	2.8–8.0	2.8–8.0	0–0.2	0–0.2	0–8.0	2.8–8.0
Min. amplitude	− 0.006	− 0.009	− 0.010	− 0.014	− 0.013	− 0.020	− 0.010	− 0.013	− 0.011	− 0.015
Max. amplitude	0.008	0.013	0.013	0.015	0.012	0.015	0.009	0.015	0.013	0.014
Pause d. (ms)[b]	∼ 500	∼ 200	∼ 200	∼ 120	∼ 120	∼ 150	∼ 100	0	∼ 100	∼ 250
"Voiced" s. (s)[c]	3	3	3.5	5	6	6	3.5	1	2	6

[a] Dominant frequency

[b] Pause duration

[c] Length of the segment that contains voice

clear and constant pauses that do not vary much in their length and frequency. This might be comparable with a sad and dragging voice or, if for example anger is expressed, simply very clear pauses. Also a more quiet speech seems to indicate a more negative emotional state. When pauses within an utterance are lasting longer, but appear more seldom over time, the network seems to have learned to predict a more positive emotional state. Moreover, the learned input signals suggest that higher frequencies in speech and longer utterances also indicate that a more extreme emotional state on valence is perceived, such as label 0 (very negative) or label 4 (very positive).

5 Discussion

In this paper, we propose a general method for learning auditory features of emotional expressions in speech. The results show that the method does not only achieve competitive classification results with state-of-the-art approaches but also enables a deeper understanding of the most relevant auditory representations in speech for emotion-related tasks.

We realize that some of the findings need to be considered with some caution. The presented results are based on the learned representations of the ELoR network, which has been trained on the annotations of one specific evaluator (*E*-2) only. Therefore, the specific results discussed in this work should not be considered universal. Nevertheless, our results confirm that the network is able to acquire phonetic characteristics that are known in linguistics to be generally relevant for arousal and valence [53].

Additionally, we emphasize that the learned representations are based on the IEMOCAP dataset only. This is a potentially limited training source for the network to completely learn the complex two-dimensional space of emotional expressions. Also, the annotations of IEMOCAP have been rated not only based on audio signals, but also on vision. Moreover, the evaluators rated the samples in a sequential order. This means that they were fully aware of the context. In our study, the proposed networks had to learn these annotations based on auditory information only. Also, the samples have been presented to the network in a random order. Therefore, it was not possible for the networks to extract any inter-context information between the samples. Thus, the network had to reproduce the same emotional states on less information than given to the evaluators of the dataset. By using a dataset that consists of annotations based on acoustic information only, the network might even learn more discriminative features. Nonetheless, the results of the conducted experiments show that the network was able to extract important and discriminative auditory representations to predict an emotional state based on sound only. Moreover, the visualizations and analysis of these learned representations give insights into the most relevant auditory features for emotional expressions in speech.

In general, we argue that an objective classification of emotional states is highly complex and difficult. The "ground truth" of an emotional state remains latent and unresolved. In emotion recognition tasks, the labels for an emotional state typically reflect the subjective perception of one or more annotators. However, this perception can differ based on the annotator, the context the annotator is aware of, the speaker, and the context the speaker is aware of. This emphasizes that emotions and their causes are complex and not fully understood. Therefore, it is important to study each of the separated domains in more detail, such as the definition of emotion, the generation of more suitable datasets, and an understanding of the most important features for each modality. This does not mean to purely focus on achieving a higher recognition performance by combining as many modalities (e.g., vision, acoustics, semantics) as possible, but to focus on understanding each modality in more depth first. Thus, the contribution of this study has not only been to propose an emotion recognition system that achieves the highest classification accuracy but also to find a general method that allows a deeper analysis and understanding of the most important features in the context of vocal emotional expressions.

5.1 Conclusions

Up to date, there is no clear consensus among researchers which auditory features are learned by deep neural networks on speech emotion recognition tasks. Thus, we propose a deep neural network topology to automatically learn features for emotional categorization of speech directly from the unprocessed signal in the time domain. Furthermore, we introduce two methods for analyzing the representations that the network learned.

As the features have been implicitly learned by the network, the learned representations are mostly independent of previous assumptions or expert knowledge in extracting features in the emotion recognition domain. The analysis has shown that the implicitly learned representations perform better than representations of handcrafted features. Moreover, the results indicate that the network has even learned new and complementary auditory features for the prediction on valence. Thus, visualization methods have been adopted to analyze the learned representations of the trained networks. The network has learned that higher frequencies, a faster-speaking rate, and longer utterances are used for a more excited emotional state. For the perception on valence, the network has learned that clear and

constant pauses indicate a more negative emotional state, while a decreasing rate of pauses, which last longer over time, tends to indicate a more positive emotional state. Moreover, a more positive or negative emotional state is predicted for utterances with higher frequencies and longer utterances. This suggests that a more extreme value on valence is correlated with high arousal. The analysis of the learned representations has also shown that a higher amplitude in a speech signal has more influence on the perceived emotional state. Also, it indicates that the prediction on arousal and on valence is based on the same speech segments.

Thus, we propose a general method that enables a deep analysis and interpretation of automatically learned representations for speech. The results show that the network was able to learn relevant representations of speech that could be interpreted in more detail. Therefore, this study contributes to a deeper understanding of the most relevant acoustic and prosodic features learned for the perception of emotional expressions in speech.

5.2 Future work

For a more general analysis of the acoustic and prosodic features that are most relevant in emotional expressions, the proposed methods can be applied to a greater amount of data. Data should be collected that includes dimensional annotations of many evaluators per recording to enable a better generalization on the perception of emotional states. For more specific results on the acoustic features, a dataset with dimensional annotations based on speech data only could be collected and studied. Also, datasets that consist of different languages could be used to compare the most relevant features in speech for different languages and cultures.

To enable a more fine-grained interpretation on the emotional expressions in speech, the proposed approaches could also be applied to a dataset that contains continuous values for valence and arousal. Then, the task could be considered as a regression problem. By adopting a loss function that is dependent on the standard deviation of the different annotations, the great variety of different perceptions of emotional states could be additionally learned by the network. Furthermore, it could be highly interesting for future research to combine the linguistic information of speech to investigate if this changes the learned representations on the acoustic information.

By using the proposed method with these suggestions, a universal understanding of acoustic and prosodic features for emotional expressions across different cultures could be gained.

Acknowledgements The authors would like to thank Florian Letsch, Tobias Hinz, Sibel Toprak, Peer Springstübe, Melanie Remmels, and Frieder Berthold for their critical comments, interesting discussions, and valuable advice during this work. The authors also gratefully acknowledge partial support from the German Research Foundation (DFG) under Project Crossmodal Learning, TRR-169.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abdel-Hamid O, Mohamed A, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. Audio Speech Lang Process 22(10):1533–1545
2. Barros P, Weber C, Wermter S (2016) Learning auditory neural representations for emotion recognition. In: Proceedings of the 2016 international joint conference on neural networks. Vancouver, pp 921–928
3. Bengio Y, Boulanger-Lewandowski N, Pascanu R (2013) Advances in optimizing recurrent networks. In: Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing, pp 8624–8628
4. Bergstra J S, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization. In: Advances in neural information processing systems, pp 2546–2554
5. Bergstra J, Yamins D, Cox DD (2013) Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms. In: Proceedings of the 12th python in science conference, pp 13–20
6. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan S (2008) IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335–359
7. Busso C, Bulut M, Narayanan S (2013) Toward effective automatic recognition systems of emotion in speech. In: Gratch J, Marsella S (eds) Social emotions in nature and artifact: emotions in human and human-computer interaction. Oxford University Press, New York, pp 110–127
8. Chang J, Scherer S (2017) Learning representations of emotional speech with deep convolutional generative adversarial networks. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2746–2750
9. Ciregan D, Meier U, Schmidhuber J (2012) Multi-column deep neural networks for image classification. In: Proceedings of the 2012 IEEE conference on computer vision and pattern recognition, pp 3642–3649
10. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human–computer interaction. IEEE Signal Process Mag 18(1):32–80
11. Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously

spoken sentences. IEEE Trans Acoust Speech Signal Process 28(4):357–366
12. Dhall A, Goecke R, Joshi J, Sikka K, Gedeon T (2014) Emotion recognition in the wild challenge 2014: baseline, data and protocol. In: Proceedings of the 16th international conference on multimodal interaction, pp 461–466
13. Dhall A, Ramana Murthy OV, Goecke R, Joshi J, Gedeon T (2015) Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 17th international conference on multimodal interaction, pp 423–426
14. Ekman P (1984) Expression and the nature of emotion. Approaches Emot 3:19–344
15. Ekman P (1992) An argument for basic emotions. Cognit Emot 6(3–4):169–200
16. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit 44(3):572–587
17. Erhan D, Bengio Y, Courville A, Vincent P (2009) Visualizing higher-layer features of a deep network. University of Montreal, Montreal, p 1341
18. Eyben F (2016) Real-time speech and music classification by large audio feature space extraction. Springer, Berlin
19. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on multimedia, pp 1459–1462
20. Fernandez R (2003) A computational model for the automatic recognition of affect in speech. Ph.D. thesis, Massachusetts Institute of Technology
21. Forman G, Scholz M (2010) Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explor. Newsl. 12(1):49–57
22. Gao Y, Li B, Wang N, Zhu T (2017) Speech emotion recognition using local and global features. In: International conference on brain informatics. Springer, pp 3–13
23. Ghosh S, Laksana E, Morency LP (2016) Representation learning for speech emotion recognition. In: Proceedings interspeech, pp 3603–3607
24. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. Aistats 9:249–256
25. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. Aistats 15:315–323
26. Golik P, Doetsch P, Ney H (2013) Cross-entropy vs. squared error training: a theoretical and experimental comparison. In: Proceedings of the 2013 Interspeech, pp 1756–1760
27. Gunes H, Pantic M (2010) Automatic, dimensional and continuous emotion recognition. Int J Synth Emot 1(1):68–99. https://doi.org/10.4018/jse.2010101605
28. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the 2015 IEEE international conference on computer vision, pp 1026–1034
29. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580
30. Huang Z, Dong M, Mao Q, Zhan Y (2014) Speech emotion recognition using CNN. In: Proceedings of the ACM international conference on multimedia, pp 801–804
31. Huang J, Li Y, Tao J, Lian Z, Niu M, Yi J (2018) Speech emotion recognition using semi-supervised learning with ladder networks. In: Proceedings Asian conference on affective computing and intelligent interaction (ACII Asia), pp 1–5
32. Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148(3):574–591
33. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167
34. Jin Q, Li C, Chen S, Wu H (2015) Speech emotion recognition with acoustic and lexical features. In: Proceedings of the 2015 IEEE international conference on acoustics, speech and signal processing, pp 4749–4753
35. Keren G, Schuller BW (2016) Convolutional RNN: an enhanced model for extracting features from sequential data. arXiv:1602.05875
36. Koolagudi SG, Rao KS (2012) Emotion recognition from speech: a review. Int J Speech Technol (Springer) 15:99–117
37. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
38. Lakomkin E, Weber C, Magg S, Wermter S (2018) Reusing neural speech representations for auditory emotion recognition. arXiv:1803.11508
39. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
40. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):541–551
41. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
42. Liu M, Chen H, Li Y, Zhang F (2015) Emotional tone-based audio continuous emotion recognition. In: International conference on multimedia modeling, pp 470–480
43. Metallinou A, Narayanan S (2013) Annotation and processing of continuous emotional attributes: challenges and opportunities. In: Proceedings of the 2013 IEEE international conference and workshops on automatic face and gesture recognition, pp 1–8
44. Muller U, Ben J, Cosatto E, Flepp B, Cun YL (2005) Off-road obstacle avoidance through end-to-end learning. In: Advances in neural information processing systems, pp 739–746
45. Nesterov Y (1983) A method of solving a convex programming problem with convergence rate O (1/k2). Soviet Math Doklady 27:372–376
46. Nummenmaa L, Saarimäki H, Glerean E, Gotsopoulos A, Jääskeläinen IP, Hari R, Sams M (2014) Emotional speech synchronizes brains across listeners and engages large-scale dynamic brain networks. NeuroImage 102:498–509
47. Patel P, Chaudhari A, Kale R, Pund M (2017) Emotion recognition from speech with gaussian mixture models and via boosted GMM. Int J Res Sci Eng 3:56–64
48. Pollack I, Rubenstein H, Horowitz A (1960) Communication of verbal modes of expression. Lang Speech 3(3):121–130
49. Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion 37:98–125
50. Ringeval F, Eyben F, Kroupi E, Yuce A, Thiran JP, Ebrahimi T, Lalanne D, Schuller BW (2015) Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognit Lett 66:22–30
51. Sainath TN, Weiss RJ, Senior A, Wilson KW, Vinyals O (2015) Learning the speech front-end with raw waveform CLDNNs. In: Proceedings of the 2015 Interspeech, pp 1–5
52. Scherer KR, Johnstone T, Klasmeyer G (2003) Vocal expression of emotion series in affective Science, handbook of affective sciences. Oxford University Press, New York, pp 433–456
53. Schröder M, Cowie R, Douglas-Cowie E, Westerdijk M, Gielen S (2001) Acoustic correlates of emotion dimensions in view of speech synthesis. In: Proceedings of the 2011 Interspeech, pp 87–90

54. Schuller BW (2013) Intelligent audio analysis. Springer, Berlin
55. Schuller BW, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun 53(9):1062–1087
56. Schuller BW, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller CA, Narayanan S et al (2010) The INTERSPEECH 2010 paralinguistic challenge. In: Proceedings of the 2010 Interspeech, pp 2795–2798
57. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. ArXiv:1312.6034
58. Song P, Zheng W (2018) Feature selection based transfer subspace learning for speech emotion recognition. IEEE Trans Affect Comput 29:32–57
59. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014) Striving for simplicity: the all convolutional net. ArXiv:1412.6806
60. Sun B, Li L, Zuo T, Chen Y, Zhou G, Wu X (2014) Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In: Proceedings of the 16th international conference on multimodal interaction, pp 481–486
61. Sutskever I (2013) Training recurrent neural networks. Ph.D. thesis, University of Toronto
62. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller BW, Zafeiriou S (2016) Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: Proceedings 41st IEEE international conference on acoustics, speech, and signal processing, ICASSP, pp 5200–5204
63. Truong KP, Leeuwen DA, Neerincx MA, Jong FM (2009) Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion. In: Proceedings of the 2009 Interspeech, pp 2027–2030
64. Weninger F, Eyben F, Schuller B, Mortillaro M, Scherer K (2013) On the acoustics of emotion in audio: what speech, music, and sound have in common. Front Emot Sci vol:4
65. Wöllmer M, Eyben F, Reiter S, Schuller BW, Cox C, Douglas-Cowie E, Cowie R et al (2008) Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In: Proceedings of the 2008 Interspeech, pp 597–600
66. Wu YT, Chen HY, Liao YH, Kuo LW, Lee CC (2017) Modeling perceivers neural-responses using lobe-dependent convolutional neural network to improve speech emotion recognition. In: Proceedings of the Interspeech, pp 3261–3265
67. Zheng WQ, Yu J, Zou Y (2015) An experimental study of speech emotion recognition based on deep convolutional neural networks. In: Proceedings of the 2015 international conference on affective computing and intelligent interaction, pp 827–831 https://doi.org/10.1109/ACII.2015.7344669