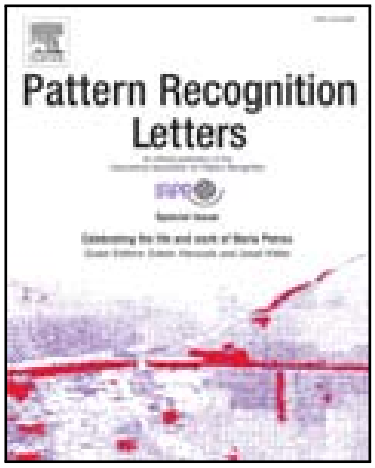
Journal Pre-proof



Speech emotion recognition via learning analogies

Stavros Ntalampiras

PII: S0167-8655(21)00031-3
DOI: https://doi.org/10.1016/j.patrec.2021.01.018
Reference: PATREC 8132

To appear in: *Pattern Recognition Letters*

Received date: 9 August 2020
Revised date: 24 November 2020
Accepted date: 17 January 2021

Please cite this article as: Stavros Ntalampiras, Speech emotion recognition via learning analogies, *Pattern Recognition Letters* (2021), doi: https://doi.org/10.1016/j.patrec.2021.01.018

Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Stavros Ntalampiras, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature Stavros Ntalampiras Date 8/7/2020

List any pre-prints: None

Relevant Conference publication(s) (submitted, accepted, or published): None

Justification for re-publication: n/a

Research Highlights (Required)

- We introduce the few-shot learning paradigm in the speech emotion recognition domain.
- Emotional characterization of speech segments is carried out by assessing analogies.
- We designed a Siamese Neural Network modeling such relationships.
- The proposed framework is able to operate in non-stationary conditions.
- Model predictions are interpretable by layer-wise investigation of the activation maps.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Speech emotion recognition via learning analogies

Stavros Ntalampiras[a,**]

[a]University of Milan, via Celoria 18, 20133, Milan, Italy

ABSTRACT

This work introduces the few-shot learning paradigm in the speech emotion recognition domain. Emotional characterization of speech segments is carried out through analogies, i.e. by assessing similarities and dissimilarities between novel and known recordings. More specifically, we designed a Siamese Neural Network modeling such relationships on the combined log-Mel and temporal modulation spectrogram space. We present thorough experimentations assessing the performance of the proposed solution holistically, where it is demonstrated that it reaches state of the art rates when following the standard leave-one-speaker-out protocol, while at the same time being able to operate in non-stationary conditions, i.e. with limited knowledge of speakers and/or emotional classes. Finally, we investigated the activation maps in a layer-wise manner in order to interpret the predictions made by the model.

1. Introduction

Affective computing including speech emotion recognition (SER) is attracting the interest of a constantly-increasing number of researchers during the last decades (Schuller, 2018; Akçay and Oğuz, 2020). Speech comprises the most relevant way of communication between humans and, in extension, is of significant importance in human computer interaction systems. As such, designing, developing and deploying emotionally-aware solutions serving applications such as smart homes, robot assistants, etc. comprises a research domain of great interest. The field of SER exploits signal processing and pattern recognition algorithms to predict the speaker's emotional state (Song and Zheng, 2018; Ntalampiras, 2020).

SER, like generalized sound and speech recognition, is based on the assumption that audio content distributions associated with different emotional states exhibit consistent patterns over time. The overall aim is to reveal and capture such distributions using suitable features and models in order to detect them in novel audio signals. In SER literature we can see the usage of a wide variety of time, frequency and wavelet domain features (Firoz Shah A. and Babu Anto P., 2017; Saste and Jagdale, 2017; Ntalampiras and Potamitis, 2014) modeled by discriminative (Lotfidereshgi and Gournay, 2017), non-discriminative (Zhiyan and Jian, 2013) as well as hybrid classifiers (Ntalampiras and Fakotakis, 2012). The vast majority of approaches are focused on a single language and only few language-agnostic methods are present in the literature, e.g. (Ntalampiras, 2020b). Moreover, it is typical that the emotional space is organized in the so-called *big-six* emotions, i.e. class dictionary $\mathcal{D} = \{angry, disgust, happy, sad, neutral, fear\}$ (Miller, 2016). In general, the literature relies on the availability of domain experts for feature/model engineering or on massive quantities of labeled data so as to address the problem in an end-to-end fashion (Tzirakis et al., 2018). Exhaustive surveys of SER literature are available in (Chandrasekar et al., 2014; Ayadi et al., 2011).

Motivated by the gaps existing in the current literature, this work introduces a scheme learning emotional speech through analogies. In other words, we transform the present classification problem into a relationship learning one. Such a scheme is able to learn similarities and dissimilarities existing in input data, and through such a functionality, carries out not only classification but addresses non-stationarities altering $\mathcal{D}$. To the best of our knowledge, there is no solution present in the literature able to consider unknown classes in their dictionary.

The present work describes the few-shot learning paradigm, where we may observe only a few samples belonging to each class before making inference(s) regarding test samples. Such

**Corresponding author: Tel.: +39-02-50316240
e-mail: stavros.ntalampiras@unimi.it (Stavros Ntalampiras)

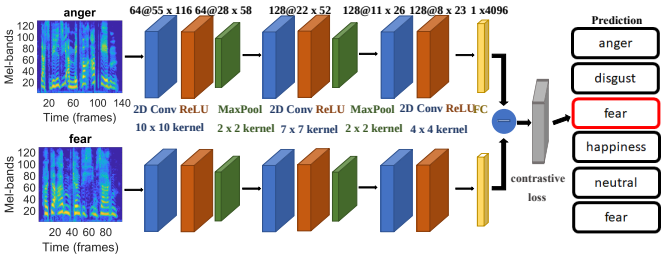


Fig. 1. The pipeline of the proposed few-speaker learning scheme using Siamese neural networks. Each input is passed though a series of convolutional, ReLU and max-pooling layers completed by a common end based on binary cross-entropy loss.

a line of thought has been explored in handwritten character recognition (Koch et al., 2015) reaching state of the art results. In the audio signal processing domain, such learning schemes remain unexplored with the exception of generative speech concepts (Lake et al., 2014).

Keeping in mind the above mentioned gaps, this work

- minimizes the need of feature engineering,
- reaches state of the art accuracy with a small amount of training data, and
- designs a reliable mechanism to detect and react to changes in $\mathcal{D}$.

More specifically, we employ two spectrogram representations emphasizing different characteristics of the audio content. Relationship learning is accomplished by means of a Siamese Neural Network composed by convolutional layers. The specific choice is motivated by the recent success of such kernels in audio pattern recognition applications (Purwins et al., 2019; Ntalampiras, 2020a), thus they could provide a solid basis for learning analogies existing between audio signals. Finally, we demonstrate the efficacy of the proposed solution via exhaustive experiments on Emo-DB dataset (Burkhardt et al., 2005) including the big-six emotional states.

In the following, we a) formalize the problem, b) delineate the proposed solution, c) describe the experimental protocol along with a detailed analysis of the obtained results, d) draw conclusions and briefly discuss potential extensions.

2. Problem formulation

In addressing speech emotion recognition, here we assume availability of a training set T^s encompassing single-channel recordings annotated according to speaker's emotional state the classes of which come from dictionary $\mathcal{D}$. Composition and cardinality of $\mathcal{D}$ are known only up to a certain extent, i.e. dictionary $\mathcal{D} = \{E_1, \ldots, E_n\}$, where E_i denotes the i-th emotional state, meaning that a-priori unknown states may appear during system's operation. At the same time, we assume that each class follows a consistent, yet unknown probability density function, which is typical for generalized audio processing systems (Ntalampiras, 2019).

Overall goal of the system is to identify speaker's emotional state in a speaker-independent manner, while at the same time being able to detect changes in composition of $\mathcal{D}$ and incorporate new appearing classes.

3. Few-shot Learning for Speech Emotion Recognition

The proposed system consists in a set of *a-priori* known vocalizations, so-called support set, and Siamese Neural Network (SNN) learning *similar* and *dissimilar* relationships of the classes in $\mathcal{D}$. The pipeline is illustrated in Fig. 1, where we see that the predicted class is the one achieving maximum similarity score. The next subsections detail the a) SNN architecture, b) feature extraction stage, and c) how the SNN carries out emotion recognition in unknown recordings.

3.1. Siamese Neural Networks

The SNN encompasses a twin network where each one processes a different input, while connected to a common ending point (Bromley et al., 1994) (see Fig. 1). There, the distance between two representations produced by each network is quantified via a predefined distance metric. Even though the networks perform their processing interdependently, their goal is to satisfy the same optimization function, hence coupling the learned weights and providing closely-located representations in the feature space. At the same time, such a learning process leads an interchangeable architecture, i.e. if the networks/inputs were to be reversed (top/bottom), the output distance value would be the same. The proposed SNN incorporates binary cross entropy loss followed by a sigmoid activation during distance assessment.

Convolutional Neural Networks (CNNs) have provided excellent performance in audio signal processing systems (Purwins et al., 2019), thus each SNN twin includes convolutional layers. Interestingly, CNNs consist in a series of stacked layers, where convolutions are succeeded by max-pooling operations. Such processing emphasizes local patterns in the 2D plane, while each hidden unit accesses only a limited part of the input, the so-called *receptive field*. Dimensionality of the learned weights is suitably controlled by max-pooling layers rendering the network indifferent to translational shifts (Piczak, 2015). It should be noted that we employed rectified linear units (ReLu), i.e. the activation function is $f(x) = max(0, x)$.

3.2. SNN architecture and learning

As shown in Fig. 1, each SNN twin encompasses three convolutional layers, where the initial two are followed by ReLu and max-pooling ones. The last layer concludes processing with a fully-connected form. SNN is completed by a distance operation succeeded by a fully connected layer and a sigmoid function assessing similarity between the elements of input's pair.

Convolutional filters have a stride equal to 1 and kernels as shown in Fig. 1, while max-pooling layers have 2×2 kernels with $stride = 2$. Learning targets the minimization of binary cross-entropy loss among network's prediction and ground truth using the standard version of backpropagation algorithm. Mini-batch size is chosen according to the TS size at a learning rate of 6e−5. Weight initialization is carried out via narrow normal

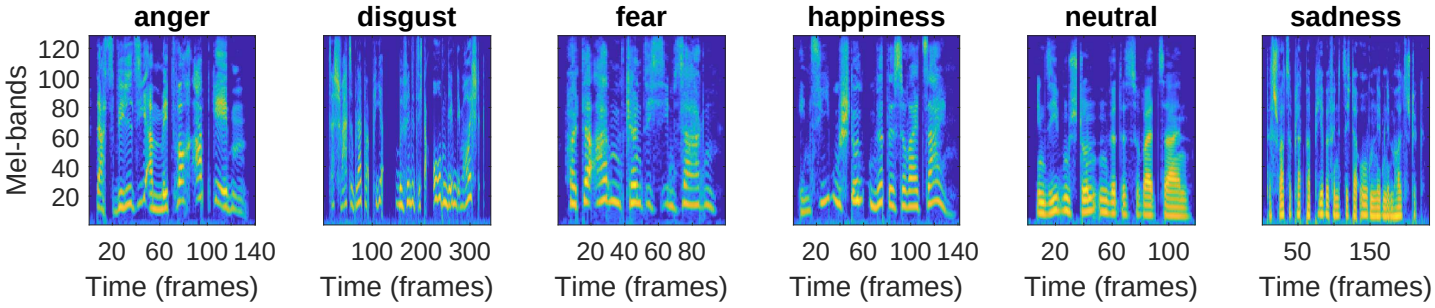


Fig. 2. Log-mel spectrograms extracted out of samples representing every available emotional state.

distributions with zero-mean and 0.01 standard deviation. The maximum number of allowed iterations is 2000.

3.3. Feature extraction

Following recent advances in affective computing (Ntalampiras, 2017), we considered two feature sets capturing different properties of the sound structure. More specifically,

Log-Mel spectrogram. the first feature set consists in Mel-scaled spectrograms representing each emotional state. We used the short time Fourier transform with size equal to 1024, while the audio signals were hamming windowed with a window size of 0.03s and 0.015s overlap. Moreover, we employed 128 equal-width log-energies following the standard Mel filter bank. Such spectrograms representing the big six emotions considered in this work are illustrated in Fig. 2.

Temporal Modulation Features. the second feature set is based on a modulation-frequency analysis conducted via Fourier transform and filtering theory as presented in (Clark and Atlas, 2009; Schimmel et al., 2007; Vinton and Atlas, 2001) using spectral center of gravity method. The main aim is to keep slow varying envelopes of spectral bands along with information regarding signals phase and structure. The algorithm assigns high values to regions of spectrum stimulating the listener's cochlea taking into account the associated temporal modulation.

Different to log-mel spectrogram, the modulation one models the human cochlea, where inner-ear vibrations are converted to electrically-encoded signals. In brief, incoming audio excites the basilar membrane which responds based on the excitation frequency. As long as such excitations differ, they stimulate unique areas of the membrane dividing cochlea's output into frequency bands. Importantly, a harmonic sound event occupying several different auditory channels exhibits analogous modulation patterns across every band. Such redundancy is the main advantage of modulation spectrogram over traditional ones when representing harmonic sounds (Klapuri, 2008).

Fig. 3 demonstrates the relationship existing between the acoustic and modulation frequency w.r.t every considered class. Temporal modulation features were extracted based on the Modulation Toolbox (Les Atlas and Schimmel, Sept. 2010).

1. Input: test speech segment v^t, trained SNN $\mathcal{N}$, dictionary $\mathcal{D}$, where each class is represented by extracted features of the support set $\langle \mathcal{S}_{i=1}^{i=|\mathcal{D}|} \rangle$;
2. Extract features f of v^t ;
3. Initialize similarity vector $V = []$;
4. **for** j=1:$|\mathcal{D}|$ **do**
   5. **for** i=1:$|\mathcal{S}|$ **do**
      6. Query $\mathcal{N}$ with the pair $\{f, \mathcal{S}_j^i\}$ and get similarity score $V(j, i)$;
   end
end
7. Predict the class maximizing the similarity score $S^* = \arg\max_{\mathcal{S}} \{V(:, i)\}$ and assign it to v^t;

Algorithm 1: The proposed speech emotion recognition algorithm based on few-shot learning ($|\bullet|$ denotes the cardinality operator).

3.4. Change Detection and Emotion Prediction

The above-described Siamese network assesses similarities/dissimilarities existing between pairs of features (log-mel and/or temporal modulation spectrogram) extracted from data available in T^s. As such, the few-shot learning paradigm can be extended to address classification tasks in a straightforward way. This is carried out by assigning the class producing the maximum similarity score to the unknown audio signal. At the same time, changes in $\mathcal{D}$ are detected when every similarity score falls below a predefined threshold. In that case, $\mathcal{D}$ is amended with a new class populated with the corresponding recording. Subsequently, the specific class is considered when processing further speech segments.

The proposed prediction algorithm, outlined in Alg. 1, extracts the spectrogram s of the unknown speech segment v^t (Alg. 1, line 2) and initializes similarity vector V (Alg. 1, line 3). Subsequently, we query $\mathcal{N}$ using the existing pair combi-

Table 1. Confusion matrix (in %) obtained with SNN trained on data coming from 9 speakers. The average recognition rate is 81.2%.

Presented \ Predicted	*Similar*	*Dissimilar*
Similar	**84.3**	15.7
Dissimilar	21.9	**78.1**

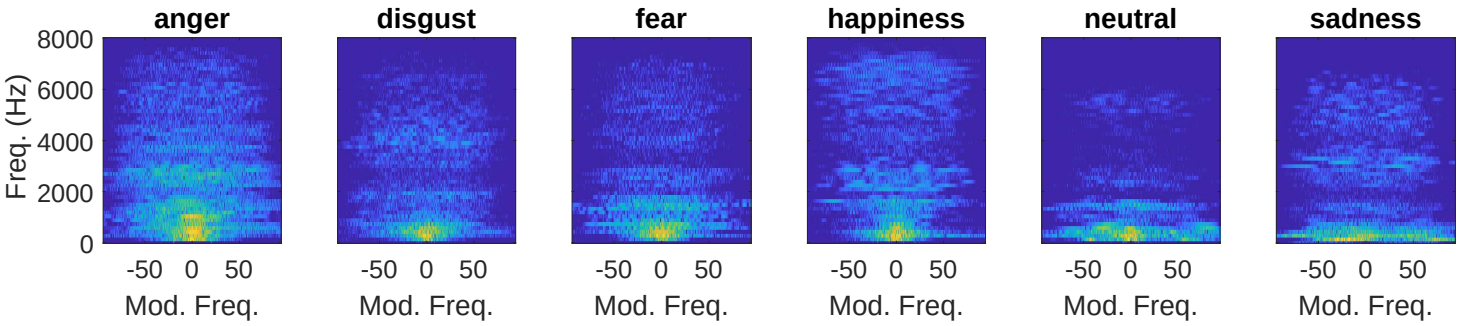


Fig. 3. Temporal modulation features extracted out of samples representing every available emotional state.

nations which outputs the corresponding similarity scores, thus updating V (Alg. 1, line 4-6). The last step of the algorithm assigns to v^t the label of the class maximizing the similarity score in V (Alg. 1, line 7).

4. Experimental set-up and results

This section describes the a) employed dataset, b) suitably formed figures of merit, c) contrasted method, d) obtained results, and e) interpretation of SNN's operation towards class assignment.

4.1. Dataset

The German language database (Emo-DB) (Burkhardt et al., 2005) encompasses voices of 10 different actors (5male/5female) expressing the following emotional states: *anger*, *disgust*, *fear*, *happiness*, *neutral*, and *sadness*. The audio is sampled at 16kHz with 16bit quantization. The distribution of samples per emotional state and actor along with age information is tabulated in Table 2.

4.2. Feature and model parameterization

To extract the feature vectors, each audio signal is framed into parts of 30ms overlapping by 15ms. The FFT size is 0.064s and the hamming window type is employed. We present results using both feature sets concurrently since their combination always outperformed their individual use. As regards to SNN, the maximum number of permitted epochs is 2000 with early stopping, the mini bath size 50, test batch 200 while the number of similarity/dissimilarity tests is 20. It should be noted that equally distributed similar and dissimilar input pairs were produced randomly. Lastly, we experimented with SNN of three and four convolutional layers; here, we report results achieved by the best-performing one, i.e. the three-layered one.

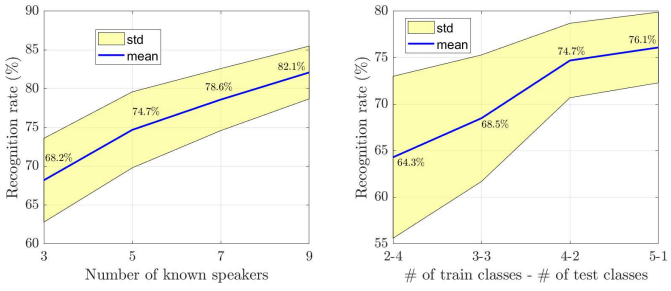


Fig. 4. Recognition rate w.r.t number of known speakers (left subfigure) and classes (right subfigure).

Table 2. Number of samples per speaker and emotional state available in EMO-DB. Speakers' sex and age are included in column S,A.

Class / *S,A*	anger	disgust	fear	happy	neutral	sad
M, 31	14	1	4	7	11	7
F, 34	12	0	6	11	10	9
F, 21	13	8	1	4	9	4
M, 32	10	1	8	4	4	3
M, 26	11	2	10	8	9	7
M, 30	12	2	6	2	4	4
F, 32	12	8	7	10	9	5
F, 35	16	8	12	8	7	10
M, 25	13	5	8	6	11	4
F, 31	14	11	7	11	5	9

4.3. Figures of merit

We employed effective and widely-used figures of merit thoroughly assessing the performance in SER. One interesting detail for the case of few-shot learning is that we can additionally employ confusion matrices demonstrating the algorithm's efficacy in recognizing similarities and dissimilarities. To this end, the following matrix was defined:

$$\mathcal{M}^s = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}, \tag{1}$$

where

- s_{11} (in %) denotes the number of times that samples fed in the first input of SNN were identified as similar to samples coming from the same class,
- s_{12} (in %) denotes the number of times that samples fed in the first input of SNN were identified as dissimilar to samples coming from the same class,

Table 3. $\mathcal{M}^s$ **(in %) achieved by SNN trained on data coming from 9 speakers (maximum rates are emboldened).**

Input \ *Output*	*Anger*	*Disgust*	*Fear*	*Happiness*	*Neutral*	*Sadness*
Anger	**86.4**	-	-	-	13.6	-
Disgust	8.3	**76.3**	-	0.5	10.8	4.1
Fear	8.8	-	**78.5**	12.7	-	-
Happiness	10.4	-	-	**77.4**	12.2	-
Neutral	4.9	-	-	10.6	**84.5**	-
Sadness	-	8.5	3.2	4	-	**84.3**

- s_{22} (in %) denotes the number of times that samples fed in the second input of SNN were identified as similar to samples coming from the same class,
- s_{21} (in %) denotes the number of times that samples fed in the second input of SNN were identified as dissimilar to samples coming from the same class.

In this case, the objective it to maximize the values appearing in the diagonal. A matrix assessing the dissimilarities $\mathcal{M}^d$ can be defined in an analogous way with the difference being that we are aiming at minimizing its diagonal. Interestingly, the sum of similarity and dissimilarity matrices characterizing the accuracy of a given method is 100%, i.e. $\mathcal{M}^s + \mathcal{M}^d = 100$ for every element.

4.4. Results

This subsection summarizes the obtained results after carrying out diverse experiments aiming at assessing different aspects of the proposed methodology. The initial phase assessed generalization over speakers. The majority of related literature follows the leave-one-speaker-out (LOSO) of T^s experimental protocol. Following the learning through analogies paradigm, we experimented with leaving several speakers out of T^s with the number of known speakers ranging from 3 to 9 (LOSO case). For each such configuration, the speakers were selected randomly, the experiment was iterated 50 times, and we report average recognition rates and standard deviation in Fig. 4a. We see that even with only 3 known speakers, the rate (68.2±5.4%) is significantly higher than chance level (16.6%).

Importantly, achieved rates increase with the number of known speakers reaching state of the art levels (Lotfidereshgi and Gournay, 2017; Chen et al., 2018) in the LOSO case, i.e. 82.1% vs. 82.8% (Chen et al., 2018). Unlike existing approaches, the proposed solution is able to operate in non-stationary environments by detecting and accommodating changes which may alter size and/or composition of $\mathcal{D}$. The corresponding matrix $\mathcal{M}^s$ is tabulated in Table 3. We observe that the class recognized best is *anger*, while *disgust* presents the highest amount of misclassifications. Such rates are in line with the ones reported in (Chen et al., 2018) where anger, sadness and neutrality are well-recognized, while the rest of classes is associated with lower recognition rates. Additionally, in the LOSO setting, we provide the confusion matrix w.r.t identification of similarities and dissimilarities in Table 1. There, we see that SNN learns similarities (84.3%) better than dissimilarities (78.1%).

The following phase evaluated generalization capabilities over emotional states. Here, only a limited amount of classes in $\mathcal{D}$ is assumed available during training, which ranges from 2 to 5. For each setting, the classes were selected randomly, the experiment iterated 50 times, and we report average recognition rates and standard deviation in Fig. 4b. Care was taken so that data coming from the same speaker is not included in both training and testing sets during the same experiment. As expected, the performance strengthens as the amount of known classes increases. We observe that the proposed solution generalizes more appropriately over speakers than classes. However, even under extremely small amount of known classes, SNN is able to learn analogies and provide rates well above chance level.

Finally, we examined the way SNN processes features by means of the considered convolutional layers emphasizing on the regions employed to assess similar/dissimilar relationships. Overall, each convolutional layer outputs a a simplified view of the obtained input image, while focusing on distinctive parts w.r.t to each class. For example, more emphasis is placed on the shape of temporal modulation spectrogram when processing *anger* w.r.t *neutral* emotional states.

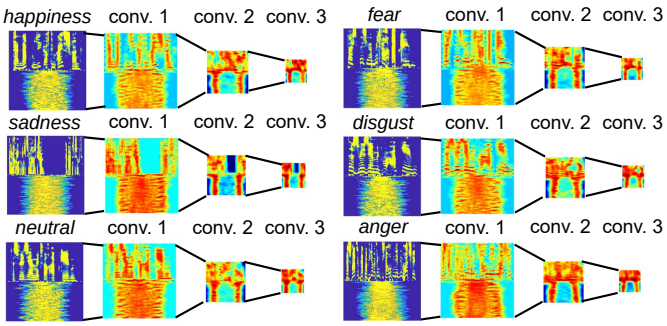


Fig. 5. Activation maps with respect to each convolutional layer for every available emotional state.

5. Conclusion

This work presented a solution learning analogies for speech emotion recognition. Importantly, current model is not specifically trained for classification but only learns similar/dissimilar relationships between input pairs. The proposed system operates via a standardized feature extraction mechanism and more importantly, is able to operate under non stationary conditions, i.e. with limited knowledge of speakers and/or class dictionary

$\mathcal{D}$. Interestingly, the incorporated change detection mechanism allows the system to react to the appearance of new emotional states and incorporate them in the class dictionary on-the-fly. It provided satisfactory performance during every experimental phase and reached state of the art accuracy in the widely used LOSO setting. At the same time, its predictions are interpretable by examining the activation maps of each convolutional layer. Large part of the success of the present learning paradigm is due to its ability to consider both similarities and dissimilarities to known classes at the same time.

In the future we are going to investigate sufficient conditions w.r.t T^s composition and quantity in order to improve the performance achieved in non-stationary environments, and extent the present framework towards transferring learned analogies to other languages.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corp. with the donation of the Titan V GPU used for this research. This work was carried out within the project entitled Advanced methods for sound and music computing funded by the Piano Sostegno alla Ricerca of University of Milan.

References

Akçay, M.B., Oğuz, K., 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication 116, 56–76. URL: https://doi.org/10.1016/j.specom.2019.12.001, doi:10.1016/j.specom.2019.12.001.

Ayadi, M.E., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44, 572–587. URL: https://doi.org/10.1016/j.patcog.2010.09.020, doi:10.1016/j.patcog.2010.09.020.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1994. Signature verification using a "siamese" time delay neural network, in: Cowan, J.D., Tesauro, G., Alspector, J. (Eds.), Advances in Neural Information Processing Systems 6. Morgan-Kaufmann, pp. 737–744.

Burkhardt, F., Paeschke, A., Rolfes, M.A., Sendlmeier, W.F., Weiss, B., 2005. A database of german emotional speech, in: INTERSPEECH.

Chandrasekar, P., Chapaneri, S., Jayaswal, D., 2014. Automatic speech emotion recognition: A survey, in: 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), pp. 341–346. doi:10.1109/CSCITA.2014.6839284.

Chen, M., He, X., Yang, J., Zhang, H., 2018. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters 25, 1440–1444. doi:10.1109/LSP.2018.2860246.

Clark, P., Atlas, L., 2009. Time-frequency coherent modulation filtering of nonstationary signals. IEEE Transactions on Signal Processing 57, 4323–4332. doi:10.1109/TSP.2009.2025107.

Firoz Shah A., Babu Anto P., 2017. Wavelet packets for speech emotion recognition, in: 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), pp. 479–481.

Klapuri, A., 2008. Multipitch analysis of polyphonic music and speech signals using an auditory model. IEEE Transactions on Audio, Speech, and Language Processing 16, 255–266. doi:10.1109/TASL.2007.908129.

Koch, G., Zemel, R., Salakhutdinov, R., 2015. Siamese neural networks for one-shot image recognition.

Lake, B.M., ying Lee, C., Glass, J.R., Tenenbaum, J., 2014. One-shot learning of generative speech concepts., in: Bello, P., Guarini, M., McShane, M., Scassellati, B. (Eds.), CogSci, cognitivesciencesociety.org.

Les Atlas, P.C., Schimmel, S., Sept. 2010. Modulation toolbox version 2.1 for matlab. URL: http://isdl.ee.washington.edu/projects/modulationtoolbox/.

Lotfidereshgi, R., Gournay, P., 2017. Biologically inspired speech emotion recognition, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5135–5139.

Miller, H.L., 2016. The SAGE Encyclopedia of Theory in Psychology. SAGE Publications, Inc. URL: https://doi.org/10.4135/9781483346274, doi:10.4135/9781483346274.

Ntalampiras, S., 2017. A transfer learning framework for predicting the emotional content of generalized sound events. The Journal of the Acoustical Society of America 141, 1694–1701. URL: https://doi.org/10.1121/1.4977749, doi:10.1121/1.4977749.

Ntalampiras, S., 2019. Generalized sound recognition in reverberant environments. JAES 67, 772–781. doi:10.17743/jaes.2019.0030.

Ntalampiras, S., 2020. Deep learning of attitude in childrens emotional speech, in: 2020 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), pp. 1–5.

Ntalampiras, S., 2020a. Emotional quantification of soundscapes by learning between samples. Multimedia Tools and Applications 79, 30387–30395. URL: https://doi.org/10.1007/s11042-020-09430-3, doi:10.1007/s11042-020-09430-3.

Ntalampiras, S., 2020b. Toward language-agnostic speech emotion recognition. Journal of the Audio Engineering Society 68, 7–13. URL: https://doi.org/10.17743/jaes.2019.0045, doi:10.17743/jaes.2019.0045.

Ntalampiras, S., Fakotakis, N., 2012. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. IEEE Transactions on Affective Computing 3, 116–125.

Ntalampiras, S., Potamitis, I., 2014. On predicting the unpleasantness level of a sound event, in: INTERSPEECH.

Piczak, K.J., 2015. Environmental sound classification with convolutional neural networks, in: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. doi:10.1109/MLSP.2015.7324337.

Purwins, H., Li, B., Virtanen, T., Schlter, J., Chang, S., Sainath, T., 2019. Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing 13, 206–219. doi:10.1109/JSTSP.2019.2908700.

Saste, S.T., Jagdale, S.M., 2017. Emotion recognition from speech using mfcc and dwt for security system, in: 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), pp. 701–704.

Schimmel, S.M., Atlas, L.E., Nie, K., 2007. Feasibility of single channel speaker separation based on modulation frequency analysis, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, pp. IV–605–IV–608. doi:10.1109/ICASSP.2007.366985.

Schuller, B.W., 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 61, 90–99. doi:10.1145/3129340.

Song, P., Zheng, W., 2018. Feature selection based transfer subspace learning for speech emotion recognition. IEEE Transactions on Affective Computing , 1–1.

Tzirakis, P., Zhang, J., Schuller, B.W., 2018. End-to-end speech emotion recognition using deep neural networks, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5089–5093.

Vinton, M.S., Atlas, L.E., 2001. Scalable and progressive audio codec, in: Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on, pp. 3277–3280 vol.5. doi:10.1109/ICASSP.2001.940358.

Zhiyan, H., Jian, W., 2013. Speech emotion recognition based on wavelet transform and improved hmm, in: 2013 25th Chinese Control and Decision Conference (CCDC), pp. 3156–3159.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.