A transfer learning framework for predicting the emotional content of generalized sound events

Stavros Ntalampiras[a)]
Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan 20133, Italy

(Received 6 October 2016; revised 10 February 2017; accepted 15 February 2017; published online 10 March 2017)

Predicting the emotions evoked by generalized sound events is a relatively recent research domain which still needs attention. In this work a framework aiming to reveal potential similarities existing during the perception of emotions evoked by sound events and songs is presented. To this end the following are proposed: (a) the usage of temporal modulation features, (b) a transfer learning module based on an echo state network, and (c) a *k*-medoids clustering algorithm predicting valence and arousal measurements associated with generalized sound events. The effectiveness of the proposed solution is demonstrated after a thoroughly designed experimental phase employing both sound and music data. The results demonstrate the importance of transfer learning in the specific field and encourage further research on approaches which manage the problem in a synergistic way. © *2017 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4977749]

[NX] Pages: 1694–1701

I. INTRODUCTION

Sound plays a fundamental role in our everyday lives carrying a great gamut of meanings and purposes, such as informative (i.e., door bell ringing), pleasant (e.g., a musical piece), alarming (e.g., a scream), relaxing (e.g., a sea wave splashing on the shore), etc. In this article, we focus on the emotional meaning conveyed by sound events. Unlike speech signals, where the speaker is able to transmit certain emotional states by altering a range of his/her vocal parameters,[1] such as fundamental frequency and loudness,[2,3] we focus on the emotion conveyed to the listener. Such sounds may be a result of human activities (e.g., walking), natural phenomena (e.g., rock falling), animals (e.g., cow mooing), etc., carrying various types of information, such as movement, size of the source, etc.[4] These may comprise the necessary stimuli for a receiver to perform various activities, for example, one may decide to take the necessary precautions in case a gunshot is heard. Such contexts demonstrate the close relationship existing between sound events and the emotions they evoke, i.e., sounds may cause emotional manifestations on the listener side, such as fear.[5]

To better understand the present problematic, one may consider the example of musical compositions where there are two types of emotions involved, i.e., the one which the composer intents to transmit and the one perceived by the audience. This study concentrates on the latter where sounds comprise a form of communication, e.g., evoking certain emotional responses (e.g., movie, radio, human computer interaction applications, etc.). In fact, the identification of the emotion on the listener's side may provide important indications towards predicting the respective human reaction.

Affective computing has received a lot of attention in the last decades with a special focus on the analysis of emotional speech, where a great gamut of generative and discriminative classifiers have been employed,[6–8] and music[9–11] where most of the literature is concentrated on regression methods. Even though generalized sound events play a major role in the emotion conveyed to the listener, they have received considerably less attention than the previously mentioned fields of research. One of the first attempts[12] considers emotions evoked by specific sounds such as a dental engine. A more generic approach[13] employed 1941 low level signal descriptors feeding a random subspace meta-learner for recognition. The authors used a dataset from FindSounds.com[14] annotated by four labelers. A well organised methodology[15] defined the structure of a sound event from the prism of the associated emotion and aimed at its automatic prediction. The authors employed a wide range of well known acoustic features along with support vector machine and artificial neural network classifiers after mapping to four categories of emotions. However, the results indicated there is no evident relationship between the waveform of a sound event and the evoked emotion(s). In the closest paper to this work,[16] the authors investigate the relationships between musical, sound, and speech emotional spaces. In particular, the authors use the characteristics of one space to predict the those of another one and vice versa, i.e., the emotion of a music piece is predicted using the speech emotional space, etc. The authors proposed a cross-domain arousal, and valence regression model showing high correlations between their predictions and the observer annotations. Their methodology is based on the INTERSPEECH 2013 Computational Paralinguistics feature set[17] feeding a support vector regression scheme.

This work is focused on the prediction of the emotional dimension of sound events, which is the area less studied in the related literature. We investigate whether the emotional space of music signals and generalized sounds is *common* since they both aim at capturing the emotions evoked to the

a)Electronic mail: stavros.ntalampiras@polimi.it

 0001-4966/2017/141(3)/1694/8/$30.00

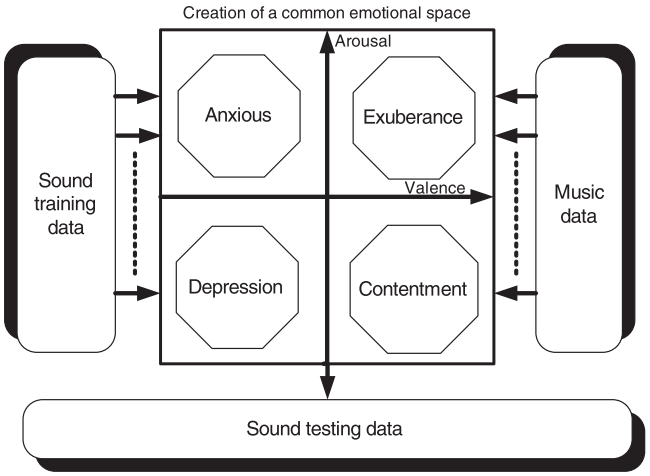

FIG. 1. The main idea behind the present work. The joint emotional space is formed by general sounds and music signals for improving the prediction of emotions evoked by sound events.

listener(s). Figure 1 demonstrates the joint emotional space composed of continuous valence and arousal values for representing the affective state of a human subject.

To assess this hypothesis we designed an experimental test bench evaluating a variety of feature sets capturing perceptual properties of the signals and several regression schemes adopted from the related literature, while we propose a version of the *k*-medoids clustering algorithm using the Minkowski distance metric. Moreover, a novel transfer learning module is incorporated for mapping the features extracted from music signals to the feature space formed by generalized sound events.

We conducted thorough experimentations on a publicly available dataset, i.e., the International Affective Digital Sounds (IADS) emotionally annotated sound events database including a diverse range of sound events (usual events, like dog barking, to uncommon ones, like gunshots or vomiting) associated with different emotional states.[18] The specific dataset is quite useful as it has been annotated following the affective annotation protocols of music emotion recognition. The music dataset is the 1000 Songs Database, which is also publicly available.[19] An extensive experimental procedure was carried out, where the superiority of the proposed method over the existing ones is demonstrated.

This work is organised as follows. Section II details the design of the proposed framework including feature extraction, transfer learning, and the *k*-medoids algorithm. Section III presents the experimental protocol starting from the specifics regarding the datasets and the parametrization of the considered solutions to the analysis of the obtained results. Finally, conclusions are drawn in Sec. IV.

II. THE PROPOSED APPROACH

We try to evaluate whether the emotional space of generalized sound events and musical pieces is shared, and if it is able to offer improved prediction of valence and arousal values. To this end the proposed framework, depicted in Fig. 2, includes the following three modules.

(a) Following the findings of our past work,[20] we exploit the temporal modulation characteristics as they have provided a performance superior to the widely used Mel frequency cepstral coefficients when applied to a task of similar perceptual needs, i.e., predicting the unpleasantness level of a sound event.
(b) Construction of the common feature space by means of transfer learning based on echo state networks (ESNs).
(c) *k*-medoids clustering algorithm.

The next three subsections provide the details regarding each module.

A. Temporal modulation features

The temporal modulation feature set is based on a modulation-frequency analysis via the Fourier transform and filtering theory.[21–23] Modulation filtering aims at retaining slow varying envelopes of spectral bands coming from non-stationary signals without affecting the signal's phase and fine-structure. This feature set assigns high frequency values to the spectrum parts affecting the cochlea of the listener while emphasizing the temporal modulation.

Unlike the power spectrogram, the modulation one originates from human cochlea modelling. There, the existing inner-ear vibration is converted to electrically encoded signals. In general, sounds excite the basilar membrane while the associated response depends on the excitation frequency. Different components must be sufficiently distinct in frequency to stimulate unique areas of the membrane, which supports the hypothesis claiming that the output of the cochlea can be divided into frequency bands. The short-time excitation energy present in a specific channel is essentially the output of the associated band. It is important to note here that a harmonic sound event occupying many different auditory channels generates a similar modulation pattern across all bands. At this point lies the basic advantage of the modulation

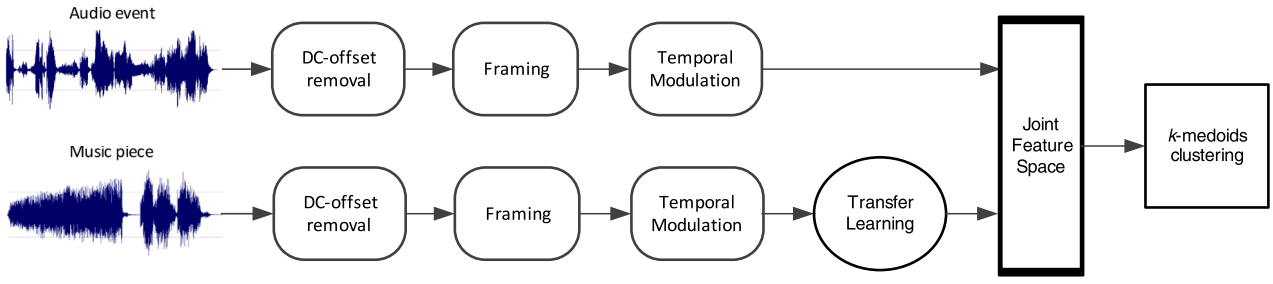

FIG. 2. (Color online) The block diagram of the proposed framework.

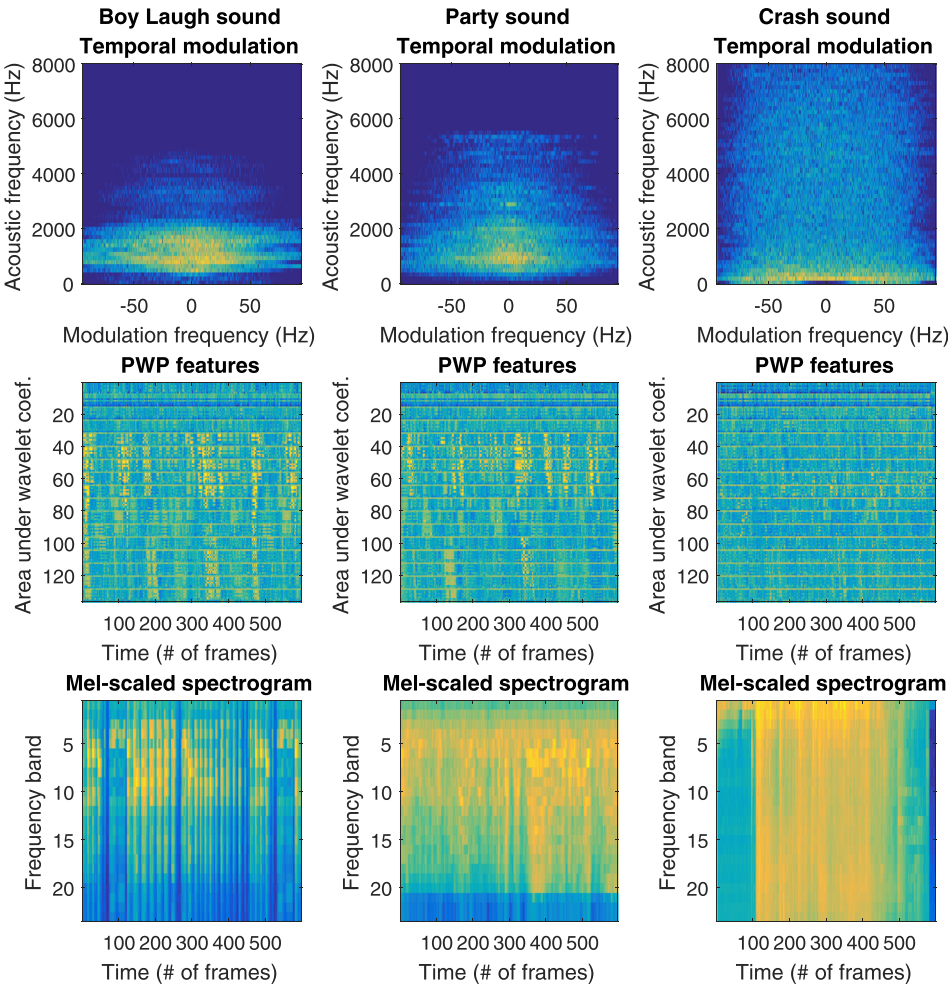

FIG. 3. (Color online) Three feature sets extracted out of sounds coming from the International Affective Digital Sounds dataset (Ref. 18). (a) Top row: temporal modulation, (b) middle row: perceptual wavelet packets, and (c) bottom row: Mel-scaled spectrograms.

spectrogram since this redundancy does not exist in conventional spectral representations of harmonic sounds.[24]

As the particular set was recently developed, a representative figure of the specific set's distribution depicting the relationship between the acoustic and modulation frequency is demonstrated in the top row of Fig. 3. It should be mentioned that the implementation of the temporal modulation features is based on the one provided in Ref. 25.

B. Transfer learning based on ESN

Feature space transformation is necessary for permitting the common handling of both feature sets by a regression methodology wishing to predict valence and arousal values of the sound events of interest. Such transformation is essential for addressing the diversities existing in the feature distributions. We overcome the particular obstacle by learning an ESN-based transformation.[26,27] It should be mentioned that this process could be performed in a vice versa manner, i.e., exploiting sound event features for characterizing music genres, which is part of our future study.

A multiple-input multiple-output (MIMO) transformation is learnt using the training data of the music and sound signals. ESN modelling, and in particular a reservoir network (RN), was employed at this stage as it is able to capture the non-linear relationships existing in the data. RNs represent a novel kind of echo-state network providing good results in several demanding applications, such as speech recognition,[27] saving energy in wireless communication,[28] etc.

An RN, the topology of which is depicted in Fig. 4, includes neurons with non-linear activation functions which are connected to the inputs (input connections) and to each other (recurrent connections). These two types of connections have randomly generated weights, which are kept fixed during both the training and operational phase. Finally, a linear function is associated with each output node.

RNs comprise a deep learning architecture as their main purpose is to capture the characteristics of high-level abstractions existing in the acquired data by designing

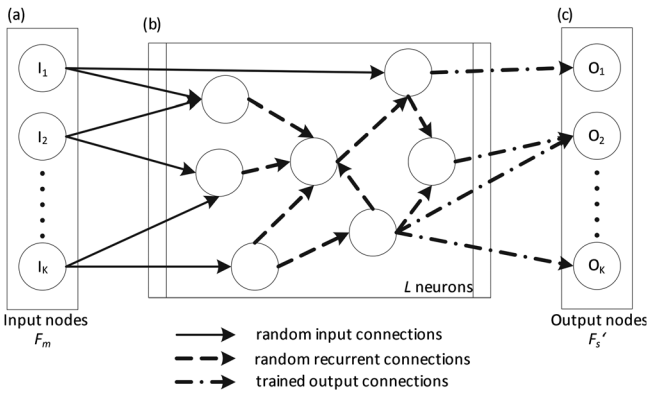

FIG. 4. The ESN used for feature space transformation.

multiple processing layers of complicated formations, i.e., non-linear functions. The associated depth is characterized by the amount of neurons included in the reservoir, which is usually called reservoir size (see Sec. III C for the parametrization analysis). Deep learning is suitable for feature space transformation facilitating the extraction of useful information for the subsequent regression modeling phase. An RN-based transformation is able to reveal and capture the highly non-linear relationships existing between the music feature space and the sound event one, which is a characteristic of significant importance facilitating the usage of a joint feature space as shown in Fig. 1.

Reservoir computing argues that since back-propagation is computationally complex but typically does not influence the internal layers severely, it may be totally excluded from the training process. On the contrary, the readout layer is a generalized linear classification/regression problem associated with low complexity. In addition, any potential network instability is avoided by enforcing a simple constraint on the random parameters of the internal layers.

In the following we explain (a) how the transfer learning *RN* (in the following denoted as *tRN*) learns the transformation from the music feature space $\mathcal{M}$ to the sound event one $\mathcal{S}$ and (b) the exact way the transformation is employed.

1. RN learning

The *tRN* is used to learn the relationships existing in the features spaces of sound and music signals. We assume that an unknown system model is followed, which may be described as a transfer function f_{RN}.

f_{RN} comprises an RN with N inputs and N outputs. Its parameters are the weights of the output connections and are trained to achieve a specific result, i.e., a sound event feature vector. The output weights are learned by means of linear regression and are called read-outs since they "read" the reservoir state.[29] As a general formulation of the RNs, depicted in Fig. 4, we assume that the network has K inputs, L neurons (usually called reservoir size), K outputs, while the matrices $W_{in}(K \times L)$, $W_{res}(L \times L)$, and $W_{out}(L \times K)$ include the connection weights. The RN system equations are the following:

$$x(k) = f_{res}(W_{in}u(k) + W_{res}x(k)), \tag{1}$$

$$y(k) = f_{out}(W_{out})x(k), \tag{2}$$

where $u(k)$, $x(k)$, and $y(k)$ denote the values of the inputs, reservoir outputs and the read-out nodes at time k, respectively. f_{res} and f_{out} are the activation functions of the reservoir and the output nodes, respectively. In this work, we consider $f_{res}(x) = \tanh(x)$ and $f_{out}(x) = x$.

Linear regression is used to determine the weights W_{out},

$$W_{out} = \arg\min_{W}\left(\frac{1}{N_{tr}}\|XW - D\|^2 + \epsilon\|W\|^2\right), \tag{3}$$

$$W_{out} = (X^TX + \epsilon I)^{-1}(X^TD), \tag{4}$$

where XW and D are the computed vectors, I a unity matrix, N_{tr} the number of the training samples, while ϵ is a regularization term.

The recurrent weights are randomly generated by a zero-mean Gaussian distribution with variance v, which essentially controls the spectral radius SR of the reservoir. The largest absolute eigenvalue of W_{res} is proportional to v and is particularly important for the dynamical behavior of the reservoir.[30] W_{in} is randomly drawn from a uniform distribution [*–InputScalingFactor*, *+InputScalingFactor*], which emphasises/deemphasises the inputs in the activation of the reservoir neurons. It is interesting to note that the significance of the specific parameter is decreased as the reservoir size increases.

Here, f_{RN} adopts the form explained in Eqs. (1), (2) by substituting $y(k)$ with F_s and $u(k)$ with F_m, where F_s denotes an original sound event feature vector and F_m a feature vector associated with a music signal.

2. Application of f_{RN}

After learning f_{RN}, it may be thought as a MIMO model of the form

$$\begin{pmatrix} F_s^{1'}(t) \\ F_s^{2'}(t) \\ \vdots \\ F_s^{N'}(t) \end{pmatrix} = f_{RN}\begin{pmatrix} F_m^{1}(t) \\ F_m^{2}(t) \\ \vdots \\ F_m^{N}(t) \end{pmatrix},$$

where the music features $F_m^1, \ldots, F_m^N$ at time t are transformed using f_{RN} to observations belonging to the sound event features $F_s^{1'}, \ldots, F_s^{N'}$, where N denotes the dimensionality of the feature vector shared by both domains. It should be noted that N depends on the feature set, i.e., temporal modulation, perceptual wavelet packets, and mel-scaled spectrum.

C. Regression based on *k*-medoids clustering algorithm

The goal of the proposed clustering algorithm is to identify feature vectors characterized by closely spaced emotional annotations. To achieve this goal we rely on the *k-mediods clustering algorithm*,[31] which belongs to the family of k-means clustering algorithms. The main characteristic of the k-mediods clustering algorithm is the ability to consider the more general concept of *pairwise dissimilarity* as distance metric, rather than the Euclidean distance (as done in the traditional k-means algorithm). This guarantees a more robust clustering phase since the effects of outliers are significantly reduced.[31]

In the general case where we have a set of elements to be clustered, the k-mediods algorithm operates as follows:[31] given the number of clusters ν to be considered (which is a user-defined parameter of the algorithm), the algorithm randomly selects ν elements from the set as medoids; a medoid is the element of a cluster whose average similarity with the rest of cluster elements is maximal. Then, the algorithm associates each element of the set to the closest medoid and

computes the overall sum of pairwise dissimilarities. This procedure, i.e., the random mediods selections and element association to each medoid, is repeated i_{max} iterations and the cluster set characterized by lowest sum of pairwise dissimilarities is chosen.

In our specific case, the elements of the k-medoids algorithm are the feature matrices. As pairwise-dissimilarity measure of the k-medoids algorithm, we propose the distance metric $d_{i,j}$ measuring the Minkowski metric between coefficient M^i and M^j defined as follows:

$$d_{i,j} = \sqrt[p]{\sum_{j=1}^{n} |M^i - M^j|^p}, \tag{5}$$

where n denotes the dimensionality of the feature vector, while p typically takes values in the interval $1 \leq p \leq 2$. Obviously the lower the values of $d_{i,j}$, the closer the respective datastreams in the Minkowski distance space. Minkowski distance was selected since it may be beneficial for tasks including responses of subjects related to a specific scale,[32] such as the one used in this work. We emphasize that $d_{i,j}$ is symmetric, i.e., and $d_{i,j} = d_{j,i}$. This process identifies the k closest matrices with respect to the unknown one M^u. Suppose that their emotional content is annotated with arousal and valence measurements are $a_1,\ldots,a_k$; $v_1,\ldots,v_k$; a_u, and v_u, respectively. The unknown characteristics of M^u are the averaged values computed as follows:

$$a_u = \frac{1}{k}\sum_{k=1}^{n} a_k \text{ and } v_u = \frac{1}{k}\sum_{k=1}^{n} v_k.$$

In the proposed sound emotion prediction system, the implementation of the k-medoids algorithm is based on partitioning around medoids.[33] Section III explains the experimental set-up and analyses the obtained results.

III. THE EXPERIMENTAL SET-UP AND RESULTS

This section explains the following.

(a) The experimental protocol that was followed towards revealing similarities between the emotions evoked by generalized sound events and music pieces.
(b) The datasets including sound and music audio signals.
(c) The parametrization of the modules included in the presented framework (see Fig. 2).
(d) The analysis of the obtained results.

A. Audio databases

For the purposes of this work, two databases have been employed.

(1) The IADS-2 (Ref. 17): This dataset includes 167 emotionally evocative sound stimuli that include contents across a wide range of semantic categories. Their annotations include two main dimensions (see also Fig. 1), i.e., valence (ranging from pleasant to unpleasant) and arousal (ranging from calm to excited). Each stimuli was rated in three separate rating studies and the final values were averaged. The selected sounds cover a broad sample of contents across the entire affective space, while they communicate emotions relatively quickly.

The subjects were female and male college students attending Introductory Psychology classes at the University of Florida. Male to female ratio was 1:1. It is important to note that each sound was rated by at least 100 participants with no hearing impairments. A preparatory phase was followed by the participants where they had to listen to three practice sounds, i.e., birds, female sigh, baby cry in order to acclimate to the types of contents that were going to be presented as well as establish the emotional rating scales.

(2) The 1000 Songs Database:[19] This dataset includes 1000 songs has been selected from the Free Music Archive.[34] Randomly (uniformly distributed) chosen excerpts with duration 45 s were subsequently isolated from each song. The songs were annotated by 100 subjects, 57 of which were males and 43 females. Their age average was 31.7 ± 10.1. A carefully designed data collection process was followed ensuring high level quality control. In fact, the participants were subjected to preliminary listening tasks, where they were asked to (a) identify music audio clips containing dynamic emotion shifts, (b) indicate the associated music genre, and (c) compile a short report explaining their willingness to address the task sufficiently as well as their competence in characterizing music content.

The annotation values are normalized in the range[1,9] facilitating transfer learning from one dataset to the other. More specifically, rates are formed such that 9 represents a high rating on each dimension (i.e., high pleasure, high arousal), and 1 represents a low rating on each dimension (i.e., low pleasure, low arousal). The normalization process aimed at the complete alignment of the annotations so that any given (pleasure or arousal) value carries exactly the same meaning amongst the datasets. It should be mentioned that no further normalization techniques were employed at this stage.

The stimulus existing in both datasets evoke reactions across the entire range of each emotional dimension, i.e., pleasure ratings for these sounds range from very unpleasant to very pleasant, and are distributed fairly evenly across the space. The observations are similar for the arousal levels as well.

B. Contrasted approaches

We compared the approach proposed here on two levels, i.e., both feature extraction and regression. More specifically we used a Mel-scaled spectrogram and the perceptual wavelet packets (PWPs) set[35] due to their ability to capture perceptual properties of audio signals. The PWP set analyses the audio signals across different spectral areas, while they are approximated by wavelet packets. They account for the fact that human perception is not affected in the same way by all parts of the spectrum[36] by employing a suitably-designed filterbank. The PWP feature set reflects upon the degree of variability exhibited by a specific wavelet coefficient within a critical band, thus they may capture useful information for characterizing emotional content. Moreover, as suggested by the related literature we compared the k-medoids clustering

algorithm with support vector regression[16] and Gaussian mixture models clustering.[20]

The contrasting experiments were designed such that all the approaches were compared on two feature spaces, i.e., the one constructed using the sound events alone and the joint one. Care was taken such that all approaches operated on identical train and test sets in order to achieve a fair comparison.

C. Parametrization of the transfer learning framework

The audio files coming from both databases were sampled at 16 kHz with 16-bit quantization and preprocessed for eliminating any possible DC-offset. The feature extraction parametrization was kept constant with respect to every set facilitating comparison and fusion tasks. After early experimentations and in order to avoid possible misalignment(s), the low-level feature extraction window is 30 ms with 20 ms overlap. Furthermore to smooth any existing discontinuities the sampled data are hamming windowed while the FFT size, where applicable, is 512.

The parameters of the *tRN* were selected by means of exhaustive search based on the minimum reconstruction error criterion. The parameters were taken from the following sets: $SR \in \{0.8, 0.9, 0.95, 0.99\}$, $L \in \{0, 500, 1000, 5000, 10000\}$, and $InputScalingFactor \in \{0.1, 0.5, 0.7, 0.95, 0.99\}$. The combination of parameters providing the lowest reconstruction error on a validation set including both feature spaces. Its implementation was based on the Echo State Network Toolbox.[37] Finally, with respect to the *k*-medoids algorithm k was set after exhaustively searching the interval[1,31] and identifying the value providing the best performance in terms of mean squared error as regards to both valence and arousal predictions.

D. Results and analysis

In the first experiment we compared the proposed solution with a plethora of approaches exploiting both different features sets and classifier for predicting the emotional properties of generalized sound events. The results are tabulated in Table I.

As we can see in Table I the best prediction results with respect to both arousal and valence measurements are provided by the temporal modulation feature set when combined with the *k*-medoids algorithm. Overall it is worth noticing that *k*-medoids based on the Minkowski metric is superior to the rest of regression approaches regardless the set of descriptors. This possibly indicates the potential ability of the temporal modulation to express characteristics associated with the emotional content of the generalized sound events. Moreover, it shows that the coefficients closely located in the feature space (in the Minkowski sense) agree on their emotional characterization. Furthermore, the poor predictions of the SVR may be due to the limited amount of data included in the IADS-2 database. This burdens the flatness of the regression function which is of fundamental importance in SVR training.[39] Moving on, we observe that GMM clustering based on the Kullback-Leibler (KL) divergence captures better the relationships existing in the feature space. Generative modelling in the stochastic plane is able to provide the second best results in arousal and valence prediction. Last it should be mentioned that the performance of the proposed approach is the best one reported on the IADS-2 dataset.

In the next experimental phase we activated the ESN-based transfer learning component and included the music data to perform prediction of the emotional content of the sound events. The parameters of the *tRN* providing the lowest reconstruction error were $SR = 0.95$, $L = 5000$, and $InputScalingFactor = 0.99$. The results are tabulated in Table II. As we can see most errors have decreased proving that (a) there exist similarities in the way song and generalized audio signals evoke emotions and (b) transfer learning for automatic description of emotional content is beneficial. More specifically, proposed method achieves MSE figures equal to 0.9 and 1.24 for arousal and valence prediction, respectively. Furthermore, the SVR approach provides lower MSEs, i.e., better performance due to the increased data availability. This is true for the GMM clustering based solution as well.

While comparing Tables I and II, the relevance of a transfer learning mechanism enabling feature space transformation becomes clear. The majority of MSEs have decreased confirming that such deep learning technique is able to transform the data successfully allowing the regressors to operate on a larger space where they are able to provide better performance.

Figure 5 depicts the influence of k on the MSE values for both arousal and valance prediction, while considering the sound event feature space alone and the joint one. It is evident that the predictions made on the joint feature space are superior to the ones made while considering the sound event space alone. As we can see the predictions are better with more neighbours in case the joint feature space is employed demonstrating that there exist strong relationships existing between the emotional content of the datasets and

TABLE I. The matrix tabulating the regression results with respect to the proposed approach and the contrasted ones while using the sound event feature space. MSE average values over 50 iterations are shown in the following format: arousal/valence, while the minimum errors, i.e., best performance, are in boldface.

Feature set \ Regressor	*k*-medoids	SVR (Ref. 16)	GMM clustering (Ref. 20)
Mel-spectrum (Ref. 38)	1.89/3.86	2.85/4.82	3.16/3.21
PWP (Ref. 35)	1.6/3.7	3.05/4.01	3.10/3.36
Temporal modulation	**1.27/2.94**	2.85/4.85	3.13/3.10

TABLE II. The matrix tabulating the regression results with respect to the proposed approach and the contrasted ones while using both feature spaces. MSE average values over 50 iterations are shown in the following format: arousal/valence, while the minimum errors, i.e., best performance, are in boldface.

Feature set \ Regressor	*k*-medoids	SVR (Ref. 16)	GMM clustering (Ref. 20)
Mel-spectrum (Ref. 38)	1.27/3.11	2.3/4.1	3.19/3.02
PWP (Ref. 35)	1.43/2.93	3/4.13	3.24/3.20
Temporal modulation	**0.91/1.24**	1.75/4.45	3.09/2.80

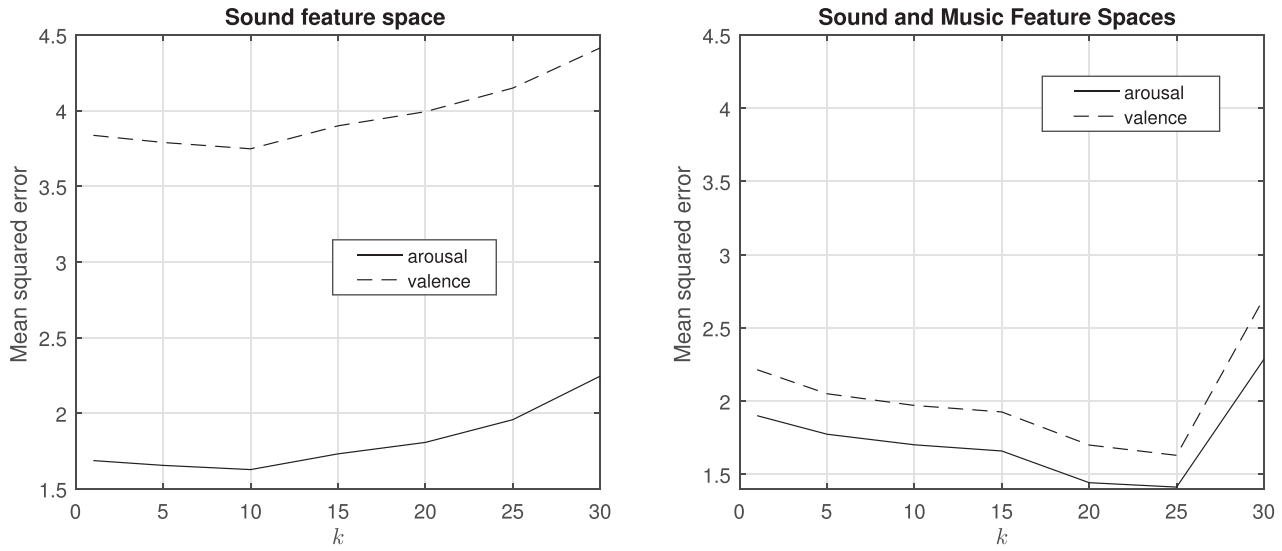

FIG. 5. The alteration of the MSE for both arousal and valance prediction as a function of k with and without including transfer learning, i.e., the song feature space.

that one can provide improved predictions using transfer learning from the musical feature space. In particular, ten neighbours provide the best performance on the space of sound events while the number increases to 25 for the joint space. In both cases, we can see that after a certain amount of neighbours, the error increases showing that the emotional content is no more relevant. An illustrative example is shown in Fig. 6 where the k-medoids algorithm identified the 25 closest neighbours to the Attack2 sound event in the joint feature space. As we can see they are composed of 19 songs and 6 sound events. The names are the ones reported in the IADS-2 and 1000 Songs databases, while the distances are analogous to the ones computed for the needs of the k-medoids algorithm.

IV. CONCLUSIONS

This paper is an attempt towards the automatic assessment of the emotions evoked by generalized sound events by revealing perceptual similarities between music and sounds via transfer learning. In particular, the presented approach proposes the usage of temporal modulation features, an ESN-based transfer learning module, and a regression solution based on the k-medoids algorithm. The proposed approach was compared with approaches exploiting Mel filterbank based and wavelet features as well as with KL divergence-based clustering and SVR. The superiority of the proposed approach was proven after a thorough evaluation employing sound and music datasets. In fact, the error rates presented here surpass the so far best published results on the IADS-2 dataset.

More importantly, we evaluated the possibility of exploiting transfer learning for constructing a shared emotional space onto which improved prediction of valence and arousal measurements was achieved. This is of critical importance and may boost further research on the specific field. The results encourage common bidirectional management of the emotion prediction domain.

Our future work includes both development of transfer learning based solutions to deal with applications of the

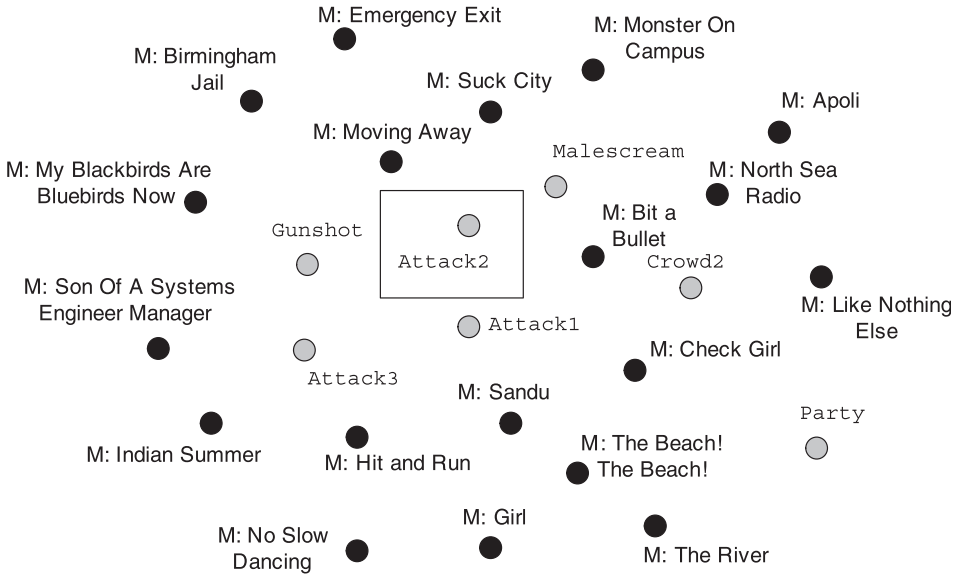

FIG. 6. The 25 nearest neighbours of the Attack2 sound event as computed by the k-medoids algorithm on the joint feature space constructed after transfer learning (a different font has been used to distinguish between sound events and songs).

generalized sound recognition technology. For example, we intent to design a synergistic framework for transferring knowledge from the music information retrieval domain to address bioacoustic signal processing applications. Finally, in the next stage of this research, a much larger dataset of sound events will be employed.

[1]Sumi Shigeno, "Effects of discrepancy between vocal emotion and the emotional meaning of speech on identifying the speaker's emotions," J. Acoust. Soc. Am. **140**, 3399–3399 (2016).
[2]Klaus R. Scherer, "Vocal communication of emotion: A review of research paradigms," Speech Commun. **40**, 227–256 (2003).
[3]V. Hozjan and Z. Kačič, "A rule-based emotion-dependent feature extraction method for emotion analysis from speech," J. Acoust. Soc. Am. **119**, 3109–3120 (2006).
[4]M. Marcell, M. Malatanos, C. Leahy, and C. Comeaux, "Identifying, rating, and remembering environmental sound events," Behav. Res. Meth. **39**, 561–569 (2007).
[5]T. Garner and M. Grimshaw, "A climate of fear: Considerations for designing a virtual acoustic ecology of fear," in *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound* (2011), pp. 31–38.
[6]M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recogn. **44**, 572–587 (2011).
[7]R. Asadi and H. Fell, "Improving the accuracy of speech emotion recognition using acoustic landmarks and Teager energy operator features," J. Acoust. Soc. Am. **137**, 2303 (2015).
[8]C. Lee, S. Lui, and C. So, "Visualization of time-varying joint development of pitch and dynamics for speech emotion recognition," J. Acoust. Soc. Am. **135** 2422 (2014).
[9]S. Fukuyama and M. Goto, "Music emotion recognition with adaptive aggregation of Gaussian process regressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), pp. 71–75.
[10]K. Markov and T. Matsui, "Music Genre and Emotion Recognition Using Gaussian Processes," IEEE Access **2**, 688–697 (2014).
[11]Y. Yi-Hsuan and H. Chen, "Machine recognition of music emotion: A review," ACM Trans. Intell. Syst. Technol. **3**, 40:1–40:30 (2012).
[12]M.-J. Gang and L. Teft, "Individual differences in heart rate responses to affective sound," Psychophysiology **12**, 423–426 (1975).
[13]B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic recognition of emotion evoked by general sound events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012), pp. 341–344.
[14]FindSounds.com, http://findsounds.com/ (Last viewed March 3, 2017).
[15]K. Drossos, A. Floros, and N.-G. Kanellopoulos, "Affective acoustic ecology: Towards emotionally enhanced sound events," in *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound* (2012), pp. 109–116.
[16]F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K.-R. Scherer, "On the acoustics of emotion in audio: What speech, music and sound have in common," Front. Psychol. **292**, 1–12 (2013).
[17]B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K-R. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *INTERSPEECH* (2013), pp. 148–152.
[18]M. Bradley and P.-J. Lang, "The international affective digitized sounds (2nd ed.; IADS-2): Affective ratings of sounds and instruction manual," technical report B-3, University of Florida, Gainesville, FL (2004).
[19]M. Soleymani, M.-N. Caro, E.-M. Schmidt, C.-Y. Sha, and Y. H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia* (2013), pp. 1–6.
[20]S. Ntalampiras and I. Potamitis, "On predicting the unpleasantness level of a sound event," in *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2014), pp. 1782–1785.
[21]P. Clark and L. Atlas, "Time-frequency coherent modulation filtering of nonstationary signals," IEEE Trans. Sign. Process. **57**, 4323–4332 (2009).
[22]S. M. Schimmel, L. E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2007), pp. 605–608.
[23]M. S. Vinton and L. E. Atlas, "Scalable and progressive audio codec," in *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)* (2001), pp. 3277–3280.
[24]A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," IEEE Trans. Audio Speech Lang. Process. **16**, 255–266 (2008).
[25]L. Atlas, P. Clark, and S. Schimmel, "Modulation Toolbox Version 2.1 for MATLAB," http://isdl.ee.washington.edu/projects/modulationtoolbox/, September 2010 (Last viewed March 3, 2017).
[26]A. Jalalvand, F. Triefenbach, D. Verstraeten, and J. Martens, "Connected digit recognition by means of reservoir computing," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association* (2011), pp. 1725–1728.
[27]D. Verstraeten, B. Schrauwen, and D. Stroobandt, "Reservoir-based techniques for speech recognition," in *International Joint Conference on Neural Networks, 2006 (IJCNN'06)* (2006), pp. 1050–1053.
[28]H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," Science **304**, 78–80 (2004).
[29]M. Lukoševičius and H. Jaeger, "Survey: Reservoir computing approaches to recurrent neural network training," Comput. Sci. Rev. **3**, 127–149 (2009).
[30]D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt, "An experimental unification of reservoir computing methods," Neural Netw. **20**, 391–403 (2007).
[31]L. Kaufman and P.-J. Rousseeuw, "Clustering by means of medoids," in *Statistical Data Analysis Based on the L1-Norm and Related Methods*, edited by Y. Dodge (North-Holland, Amsterdam, 1987), pp. 405–416.
[32]J. P. van de Geer, "Some aspects of Minkowski distance," research report, Department of Data Theory, University of Leiden (1995).
[33]S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd ed. (Academic, Orlando, FL, 2006), pp. 1–856.
[34]Free Music Archive, http://freemusicarchive.org/ (Last viewed March 3, 2017).
[35]S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Exploiting temporal feature integration for generalized sound recognition," EURASIP J. Adv. Sig. Proc. **2009**, 807162 (2009).
[36]B. Scharf, "Complex sounds and critical bands," Psychol. Bull. **58**, 205–217 (1961).
[37]Echo State Network Toolbox, http://reservoir-computing.org/software (Last viewed March 3, 2017).
[38]L. Yi-Lin and W. Gang, "Speech emotion recognition based on HMM and SVM," Int. Conf. Mach. Learn. Cybern. **8**, 4898–4901 (2005).
[39]A.-J. Smola and B. Schölkopf, "A tutorial on support vector regression," Stat. Comput. **14**, 199–222 (2004).