Transfer learning emotion manifestation across music and speech

Eduardo Coutinho, Jun Deng, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Coutinho, Eduardo, Jun Deng, and Björn Schuller. 2014. "Transfer learning emotion manifestation across music and speech." In *2014 International Joint Conference on Neural Networks (IJCNN), 6-11 July 2014, Beijing, China*, 3592–98. Piscataway, NJ: IEEE. https://doi.org/10.1109/ijcnn.2014.6889814.

Transfer Learning Emotion Manifestation Across Music and Speech

Eduardo Coutinho, Jun Deng, and Björn Schuller

Abstract—**In this article, we focus on time-continuous predictions of emotion in music and speech, and the transfer of learning from one domain to the other. First, we compare the use of Recurrent Neural Networks (RNN) with standard hidden units (Simple Recurrent Network - SRN) and Long-Short Term Memory (LSTM) blocks for intra-domain acoustic emotion recognition. We show that LSTM networks outperform SRN, and we explain, in average, 74%/59% (music) and 42%/29% (speech) of the variance in Arousal/Valence. Next, we evaluate whether cross-domain predictions of emotion are a viable option for acoustic emotion recognition, and we test the use of Transfer Learning (TL) for feature space adaptation. In average, our models are able to explain 70%/43% (music) and 28%/11% (speech) of the variance in Arousal/Valence. Overall, results indicate a good cross-domain generalization performance, particularly for the model trained on speech and tested on music without pre-encoding of the input features. To our best knowledge, this is the first demonstration of cross-modal time-continuous predictions of emotion in the acoustic domain.**

I. Introduction

Speech prosody is the pattern of acoustic changes within spoken utterances that communicate meaning, including emotional, independently of verbal understanding. The acoustic changes occur as modulations of speed and continuity, accentuation, pitch and range, timbre and dynamics of speech and vocalizations. Music, like speech, has the capacity to communicate emotions to listeners through the organization of acoustic signals (e.g., [1]). As in the case of speech, listeners construe emotional meaning by attending to structural aspects of the acoustic signal and there is evidence of specific acoustic cues and patterns communicating similar emotions to listeners (see [2]).

Given the striking similarities between both domains, emotion psychologists and computer scientists ([3], [4], [5], [6]) have recently begun to compare the acoustic cues to emotion in music and speech. These studies provide evidence of the existence of acoustic profiles common to the expression of emotion in speech and music, with particular acoustic codes consistently associated with particular emotions. Give this context, the main focus of this paper is to explore the similarities between the acoustic structure of music and speech signals in the communication of emotion. From a machine learning perspective it is advantageous to explore such a relationship since the undifferentiated use of music and speech signals can enlarge the amount of available data which can be used to improve the performance of acoustic emotion recognition systems in both domains. Furthermore, it can lead to the development of hybrid systems, i.e., applicable to both music and speech signals. To this aim we will evaluate cross-modal predictions of emotion in speech and music, and evaluate the use of TL techniques to deal with differences in the feature space and data sets distributions of both types of stimuli.

Following the findings in [7], [8] and [9], Coutinho and Dibben ([5]) have shown that emotion in music and speech can be predicted from spatio-temporal patterns of low-level acoustics using SRN. Such findings, highlight the relevance of time-continuous models sensitive to the temporal context of acoustic cues to the prediction of emotion and the need to continuously adapt the predictions of emotion qualities, which is particularly important given the dynamic nature of emotional expression in both music and speech (i.e., communicated affect can vary within a music piece or a sentence). Moreover, for the application of acoustic emotion recognition to naturalistic settings, where continuous streams of information are always available, predicting emotion in a time-continuous is essential to improve interactions between humans and machines. For these reasons, this paper deals with time-continuous predictions of emotional responses to music and speech and we will recur to modelling paradigms sensitive to the temporal context of acoustic cues.

II. Related Work

Only a few works have addressed the issue of time-continuous predictions of emotion in speech and music signals. In [7] and [8] the authors used SRNs to model time-continuous emotional responses to Western Art music. In [10], the authors have replicated their method with a new set of music pieces, and extended their model to incorporate physiological cues as predictors of experienced emotions. In [5] the same methodology as been used to model time-continuous emotional responses to a film music as well as natural speech. In these studies, a large group of human annotators was used (between 40 and 60) and the data was psychometrically reliably, i.e., time-continuous annotations were highly consistent across subjects. [11] used LSTM networks to predict ratings of Arousal and Valence from a subset of natural speech recordings from the SAL Database ([12]). In [13] a similar method was used to predict continuous emotion dimensions from speech signals. Both works recur to only 4 annotators. Finally, the 2012 Audio-Visual Emotion Challenge (AVEC 2012) dedicated one sub-challenge to time-continuous predictions of emotion from audio (and video) ([14]).

E. Coutinho (phone: +49 (0)89 289-28549; fax: +49 (0)89 289-28535; e-mail: e.coutinho@tum.de), J. Deng and B. Schuller are with the Machine Intelligence & Signal Processing Group, Institute for Human-Machine Communication, Technische Universität München, Germany.

E. Coutinho is also with the School of Music, University of Liverpool, United Kingdom (e-mail: e.coutinho@liv.ac.uk). B. Schuller is also with the Department of Computing, Imperial College London, United Kingdom (e-mail: bjoern.schuller@imperial.ac.uk).

This work was supported in part by the European Research Council in the European Community's 7th Framework Program under grant agreements No. 338164 (Starting Grant iHEARu) and 230331 (Advanced Grant PROPEREMO).

In this paper, we recur to data set published in [5] and extend the authors' work by adopting a new modeling strategy based on LSTM networks. The rationale behind this decision is that LSTM networks have shown remarkable performance in a variety of pattern recognition tasks, and consistently outperform SRNs. Furthermore it has been successfully used in a similar machine learning context for speech emotion recognition (e.g., [11]). Therefore, we propose to combine the best of the various approaches previously described: a psychometrically reliable "ground truth", and state-of-the-art context dependent neural networks to model spatiotemporal acoustic patterns. As in [5], we will focus on a small set of theoretically chosen acoustic features which has the advantage of largely reducing computation times (both in training and performance).

III. Methodology

A. Acoustic features to emotion in speech and music

As mentioned in the introduction, in [5] it has shown that emotion in speech and music can be predicted to a great extent from a set of only seven (psycho)acoustic features. In this work, we will use the same set of seven features, plus a measure of auditory dissonance, that belong to six psychoacoustic dimensions consistently associated with the communication of emotion in the acoustic domain (summary of features also in Table I):

- **Intensity**: Loudness is the perceptual correlate of sound intensity (or physical strength) which we quantified using Chalupper and Fastl's dynamic loudness model ([15]).
- **Duration** The measures of duration consist of the rate of speech and musical tempo. The former was estimated using De Jong and Wempe's algorithm ([16]), which detects syllable nuclei and quantifies speech rate as the number of syllables per minute (SPM). The latter was estimated from the inter-beat intervals obtained for each piece using BeatRoot ([17]), and quantified as the number of beats per minute (BPM).
- **Pitch** The perceived pitch level and pitch contour were calculated separately for music and speech. The prosodic contour was calculated using Prosogram ([18]), a prosody transcription tool that estimates the intonation contour (the perceptual correlate of the fundamental frequency, F0) as it is perceived by human listeners. The melodic contour was estimated using a toolbox for automatic transcription of polyphonic music ([19]). In addition to these measures we also calculated the spectrum flux for all stimuli in order to quantify how much the power spectrum of the signal changes in time.
- **Timbre** Timbre was quantified using: 1) The power spectrum centroid which is calculated by the weighted mean of the frequencies present in the signal (a quantity strongly associated with the impression of sound "brightness"); and 2) A sharpness measure proposed by Aures ([20]) that approximates the subjective experience of sharpness on a scale ranging from dull to sharp (measured in acum).
- **Roughness** The term auditory roughness describes the perceptual quality of buzz, raspiness or harshness associated with narrow harmonic intervals, and is a perceived correlate of dissonance. We quantify Roughness using Vassilakis ([21]) algorithm and auditory dissonance using Aures ([20]) formula.

TABLE I. Acoustic and Psychoacoustic features used in this study

Feature	Definition	Domain
Loudness	Perceptual correlate of sound intensity	M, S
Tempo	Beats per minute (BPM)	M
Speech Rate	Number of syllables per minute (SPM)	S
Melody Contour	Salient stream of audible pitches from the full harmonic structure of the polyphonic signal	M
Prosody Contour	The perceptual correlate of the fundamental frequency (F0)	S
Spectral Flux	Amount of change in the power spectrum of the signal over time	M, S
Sharpness	The subjective experience of sharpness on a scale ranging from dull to sharp	M, S
Spectral Centroid	The weighted mean of the frequencies present in the signal	M, S
Roughness	Perceptual quality of buzz, raspiness or harshness	M, S
Auditory dissonance	Sum of the energy of the beating frequencies in auditory channels	M, S

Note: M = Music; S = Speech

B. Data sets

For this study we used the data described in [5], which was obtained in a controlled empirical study focused on the expression of emotion in real-life scenarios. The music set consisted of 8 unaltered pieces of film music summing up to approximately 14 minutes (ranging from 84s to 130s), obtained from late 20^th century Hollywood film scores. The speech set contained 9 natural speech samples (circa 13 minutes in total, ranging from 45s to 156s) by 8 different speakers, excerpted from commercially available and online films, dramatic performances, poetry recitations and TV interviews. All samples were chosen to be from the same language (German) not understood by the rather. This was done in order to avoid any confounds related to the semantic content of speech. More details about the dataset can be found in [5].

Emotions were quantified using a two-dimensional model of affect consisting of Arousal and Valence dimensions ([22]). In this model, emotions are represented as points in a two dimensional space, whose location determines their experiential qualities in terms of relative Valence (ranging from positive to negative affect) and Arousal (ranging from high to low neurophysiological alertness). Raters used a computer mouse to control a cursor on the screen whose position indicates the level of Arousal and Valence perceived at each moment in the music or speech sample. The vertical axis represented Arousal (ranging from low - bottom of the screen - to high - top of the screen), and the horizontal one Valence (ranging from very unpleasant - left side of the screen - to very pleasant - right side of the screen). Values were recorded every time the mouse was moved with

a precision of 1 millisecond. In order to eliminate linear offsets, the time series resultant of the ratings of the 52 annotators were (on an rater basis) standardized to zero mean and unit standard deviation. Then, the standardized individual time series were averaged across all oracles to obtain a pair of time series depicting the collective time-continuous annotations of Arousal and Valence for each instance. Finally, the time series correspondents to each music piece and speech sample were resampled from the original sample rate of 1000Hz to 1Hz since, typically, no relevant changes in annotations occur at faster rates (see [23]). This data is the "ground-truth" used in this experiments reported in this paper. Table II summarizes the details of the database.

TABLE II. DATABASE DETAILS

Number of instances	8	9
Time frame length	1s	1s
Average number of time frames per instance	105	80
Total number of time frames	845	723
Number of Speakers	-	8
Gender of speakers	-	4 female 4 male

C. Long-short term memory networks

In our experiments, we consider the contribution of an advanced technique for neural network based context modeling: Long Short-Term Memory (LSTM) [24]. LSTM networks make use of special memory blocks, rather than the conventional hidden cells used in typical neural networks, which endow the model with the capacity of accessing a long-range temporal context and predicting the outputs based on such information. Such memory blocks overcome a major problem of traditional recurrent neural networks (RNN) whereby the the influence of the network inputs on the hidden units (and therefore the outputs) decays or blows up exponentially as the information cycles through the network recurrent connections.

An LSTM network is similar of a simple RNN except that the nonlinear hidden units are replaced by a special kind of memory blocks. Each memory block comprises one or more self-connected memory cells and three multiplicative units – input, output and forget gates – which provide the cells with analogues of write, read and reset operations. The multiplicative gates allow LSTM memory cells to store and access information over long sequences (and corresponding periods of time) which permits to overcome the vanishing gradient problem of simple RNN. Fig. 1 shows a LSTM memory block with a one cell.

The cell input is first scaled by the activation of the input gate. Then the output is computed based the activation of the output gate, plus the memory cell values in the previous time step, which is controlled by the activation of the forget gates. For a particular the memory block, if $\boldsymbol{W}$ is a weight matrix, $\boldsymbol{x}_t$ is the input vector, $\boldsymbol{h}_t$ is the hidden vector, $\boldsymbol{b}_t$ is the hidden bias vector, then the activation vector of input gate $\boldsymbol{i}_t$ can be expressed as follows:

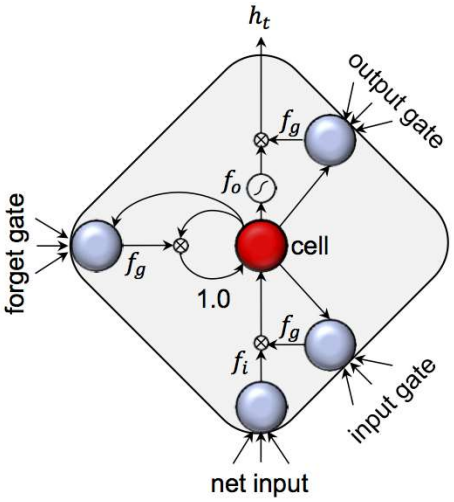

Fig. 1. LSTM memory block

$$i_t = f_g(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i), \quad (1)$$

where, $\boldsymbol{f}_g$ denotes the logistic sigmoid function.

Similarly, the activation of the forget gate $\boldsymbol{f}_t$ can be written as

$$f_t = f_g(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f). \quad (2)$$

The memory cell value $\boldsymbol{c}_t$ is the sum of input at time t and its activation in the previous time step:

$$c_t = f_i(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) + \boldsymbol{f}_t \cdot \boldsymbol{c}_{t-1}, \quad (3)$$

where f_i is the tanh activation function.

The output of the memory cell values is controlled by the output gate activation of

$$o_t = f_g(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_t + \boldsymbol{b}_o), \quad (4)$$

and the final output of the memory block is

$$h_t = \boldsymbol{o}_t \cdot f_o(\boldsymbol{c}_t), \quad (5)$$

where f_o is also a tanh activation function.

LSTM networks have shown remarkable performance in a variety of pattern recognition tasks, including phoneme classification [25], handwriting recognition [26], keyword spotting [27], and driver distraction detection [28]. Furthermore they have also been used on the context of continuous measurements and acoustic emotion recognition (e.g., REF CHINA PAPER). For more details on LSTM networks please refer to [24].

D. Transfer Learning and Auto-Encoder Pre-Training

TL has its origins on the observation that humans can apply previous knowledge to solve new problems faster and intelligently, and it has been proposed to deal in convenient ways with the problematic situation of having training and test samples derived from different feature spaces and distributions. Indeed, when the distribution of the data changes,

most statistical models need to be rebuilt using new training data, which is expensive and often impossible. In the speech emotion recognition field, a common limitation of emotion recognition models previously trained on a specific speech corpus is the tendency to perform better to that same corpus than to data retrieved from other sources. This is due to the characteristics of the speakers in each corpus, as well as the type of emotions being conveyed, the level of portrayal vs. spontaneity, among others (see [29]). Because of this, TL has recently started to be applied to this area due to its multiple advantages in dealing with mismatches between training and test data sets ([30]).

In the context of the work presented in this paper, and particularly to the aim of demonstrating the existence of shared acoustic codes communicating emotions in music and speech, we face a different challenge: how to allow a model trained on music signals be used to predict emotions expressed in the voice (and vice-versa) given differences in the feature space and distributions of both types of stimuli? To address this issue we propose the application of representation learning techniques ([31]) in order to learn transformations of the data that ease the extraction of useful information when building the regression models proposed. In particular, we implemented a denoising autoencoder (DAE), a more recent variant of the basic autoencoder consisting of only one hidden layer, that is trained to reconstruct a clean "repaired" input from its corrupted version [32]. In so doing, the learner must capture the structure of the input distribution in order to reduce the effect of the corruption process [32]. This method has been shown to be a simple but well-suited solution in previous work (e.g., [33]), including acoustic emotion recognition (e.g., [30]).

IV. Experimental Setup

A. Models training

We conducted four separate modeling experiments. In the first two experiments, we focused on intra-domain predictions and trained separate models to predict emotional responses to music pieces and speech samples. Given that we have 8 full music pieces and 9 different speech instances in our database, we used an 8-fold cross-validation procedure for the music model (each of the 8 fold consisting uses 7 pieces for the development/training set and the remaining one as test set), and a 9-fold one for the speech model (similar cross-validation process).

In the remaining two experiments we focused on cross-domain predictions. In one, we trained a model on the full music set and evaluated its performance on the full set of speech instances, and vice-versa, without pre-transformation of the features. In another, we previously adapted the input space of each by training a DAE to transform the acoustic features of music and speech instances into common representations.

Each DAE was trained to reproduce the feature space of acoustic descriptors of the full set of music pieces (Music to Speech) or speech samples (Speech to Music). During the DAE learning process, we applied the stochastic gradient descend method with momentum to optimize the parameters of the DAE. The number of hidden units was fixed to 8 (the same number of acoustic features used to encode music and speech) and attempted weight decays values were the following: $\lambda \in \{1e-4, 1e-3, 1e-2, 1e-1\}$. Additionally, masking noise with a variance of 0.2 was injected in the input layer of the DAEs. After training, the hidden layer activations of each DAE were computed for both music and speech data sets and replaced the original input features. As in the "direct" cross-domain experiments, one model was trained on the full set of music (or speech) and tested on the full set of speech (or music), but this time using the transformed features obtained from the DAE.

In all four experiments, we computed 20 trials for each model, all with randomized initial weights in the range [-0.1,0.1]. After preliminary tests, we settled with an architecture consisting of 2 hidden layers, each with 12 hidden units (with tangent activation functions in the case of the SRN, and LSTM memory blocks in the case of the LSTM network). The learning rate for all models and trials was 0.001 and momentum 0.9. An early stopping strategy was used to avoid overfitting the training data. Training was stopped after 20 iterations without improvement to the performance of the test set, and a maximum of 2000 total iterations of the learning algorithm was allowed. Each sequence (music piece or speech sample) in the various training sets was presented randomly to the model during training. The input (acoustic features) and output (emotion features) data was standardized to the mean and variance of the correspondent training sets used in each experiment and cross-validation fold of the intra-domain models.

V. Results

Four measures are used to quantify the models' performance: two indexes of precision - the root mean squared error ($rmse$) and its normalized version - $nrmse$ (allows to infer the magnitude of the $rmse$) by normalizing it to the range of observed values), and a measure of similarity - Pearson's linear correlation coefficient (r) - and its squared value (R^2; explained variance). For the performance calculations, the time-continuous outputs averaged across the 5 best trials (those with the lowest rmse for both outputs) of each experiment was used. All measures were calculated separately for the test set of each cross-validation fold, and then averaged across all folds (therefore they estimate the performance of the models for novel instances only).

A. Intra-domain experiments: Music and Speech

In the intra-domain experiments we evaluated the extent to which time-continuous responses to music and speech could be predicted using a set of training examples belonging to the same acoustic domain. For both domains, we compared the performance of SRN and LSTM networks. The results obtained in the intra-domain experiments are shown in Table III.

Both Music and Speech models exhibit a good performance using both neural network types (SRN and LSTM). Indeed, the maximum rmse corresponds to 15% of the total range of predicted values. It is clear nevertheless, that the LSTM network performs better than the SRN, both in term of precision as of similarity and explained variance (see values in bold in Table III). The only exceptions the Valence

TABLE III. INTRA-DOMAIN PERFORMANCE RESULTS: $rmse$, $nrmse$, r AND R^2.

Performance Measure	Variable	Music (SRN)	Music (LSTM)	Speech (SRN)	Speech (LSTM)
rmse	Arousal	0.34	**0.28**	0.39	**0.38**
	Valence	0.39	**0.34**	0.37	**0.34**
nrmse	Arousal	12 %	**10 %**	**11 %**	**11 %**
	Valence	15 %	**13 %**	12 %	**10 %**
r	Arousal	0.71	**0.82**	0.42	**0.59**
	Valence	0.25	**0.51**	**0.29**	0.27
R2	Arousal	59 %	**74 %**	26 %	**42 %**
	Valence	29 %	**59 %**	**29 %**	**29 %**

TABLE IV. CROSS-DOMAIN PERFORMANCE RESULTS: $rmse$, $nrmse$, r AND R^2.

Performance Measure	Variable	S ⇒ M	S ⇒ M (TL)	M ⇒ S	M ⇒ S (TL)
rmse	Arousal	**0.52**	0.60	0.39	**0.38**
	Valence	**0.71**	0.72	0.37	**0.34**
nrmse	Arousal	**19 %**	22 %	18 %	**16 %**
	Valence	**26 %**	27 %	18 %	**16 %**
r	Arousal	**0.80**	0.68	0.05	**0.32**
	Valence	**0.14**	0.09	0.06	**0.16**
R2	Arousal	**70 %**	54 %	12 %	**28 %**
	Valence	**43 %**	34 %	7 %	**11 %**

output of the Speech model, whose performance does not differ from the SRN.

By analyzing in more detail the performance of the LSTM models, we can observe that both the Music and Speech models exhibit a low prediction error, ranging from 10% to 13% of the total output range. Nevertheless, clear differences emerge when looking at the coefficients r and R2. In this regard, the Music model makes better predictions and explains more variance compared to the Speech one. This is true for both Arousal and Valence. The Music model is able to explain an average of 74% of the variance in Arousal (vs. 42% of the Speech model) and 59% of the variance in Valence (vs. 29% of the Speech model).

B. Music to Speech and Speech to Music

Given that in the previous section the LSTM network performed better than the SRN, in these experiment we focus on using the former. In Table IV we show the results of the LSTM-based models trained on the Speech instances and tested on Music ones (S⇒M) and vice versa (M⇒S). In both cases, we show the results with and without the use of TL, i.e., using or not a DAE to adapt the input feature space. The goal is to establish whether feature domain adaptation is advantageous to cross-modal applications.

Looking first at the S⇒M model, we see that the simplest version, i.e., the one without the DAE, performed better. This is obvious in all performance measures. This model, trained on speech, explains 70% of the variance in Arousal and 43% in Valence. Compared to the correspondent intra-domain Music model (see Table III, it is only 4% lower for Arousal and 16% for Valence. The nrmse is nevertheless approximately the double for both outputs, thus indicating lower precision.

In relation to the M⇒S experiment, the model with the DAE performed better. Again this is also visible in all performance measures. This model, trained on music, explains 28% and 11% of the variance in, respectively, Arousal and Valence. Contrasted to the correspondent intra-domain Speech model (see Table III), it is 14% lower for Arousal and 18% lower for Valence. Contrary to the S⇒M model, the nrmse is only 5% higher for Arousal and 6%. Thus, precision figures are much closer.

Figures 2 and 3 depict the target annotations versus the best intra- and cross-domain models' predictions for 4 instances of each domain. Performance measures are also indicated for each instance.

VI. DISCUSSION AND CONCLUSIONS

In this article, we have have focused on time-continuous predictions of emotion in music and speech, and the transfer of learning from one domain to the other. We started by reproducing the work in [5] by implementing a SRN for the prediction of emotion in music and speech from a small set of acoustic features. Next, we developed further the method and adopted a new modeling paradigm based on LSTM networks, and, as expected, we have shown that LSTM networks are a viable and profitable alternative to SRNs for predicting time-continuous emotions from acoustic signals. It is worth noticing that we have achieved a very good performance with only a set of 8 acoustic features. It remains to be evaluated in future work whether such a minimalistic approach is a viable alternative to the use of large feature spaces (hundreds or thousands of features).

Our next major goal was to evaluate whether cross-domain predictions of emotion are a viable option for acoustic emotion recognition, considering the close link between music and speech in the communication of emotion (see [3]), and a recent successful demonstration in the discrete domain ([6]). To this end, we also evaluated the use of a standard TL technique for input feature space adaptation (a denonising auto-encoder). Overall, results indicate a good cross-domain generalization performance. This was particularly evident for the model trained on Speech and tested on Music without TL pre-encoding of the input features.

The fact that we obtained a better performance from speech to music can simply reflect the nature of the dataset used, but it raises some issues in relation to the symmetry of the knowledge transfer. In either case, further tests with other databases (ideally larger) are necessary to clarify this issue. In relation to the use of the proposed DAE, results are contradictory as the M⇒S performed better with with the input features space transformation but the S⇒M did not. More tests are needed in order to establish the need to adapt the features space of source and target data. To our best knowledge, this is the first demonstration of cross-modal time-continuous predictions of emotion in the acoustic domain. Nevertheless, due to the small size of the data sets it remains to be determined the extent to which our results can be applicable to a wider range music pieces and speech samples.

REFERENCES

[1] P. N. Juslin and J. A. Sloboda, *Handbook of music and emotion: Theory, research, applications.* Oxford University Press, 2010.

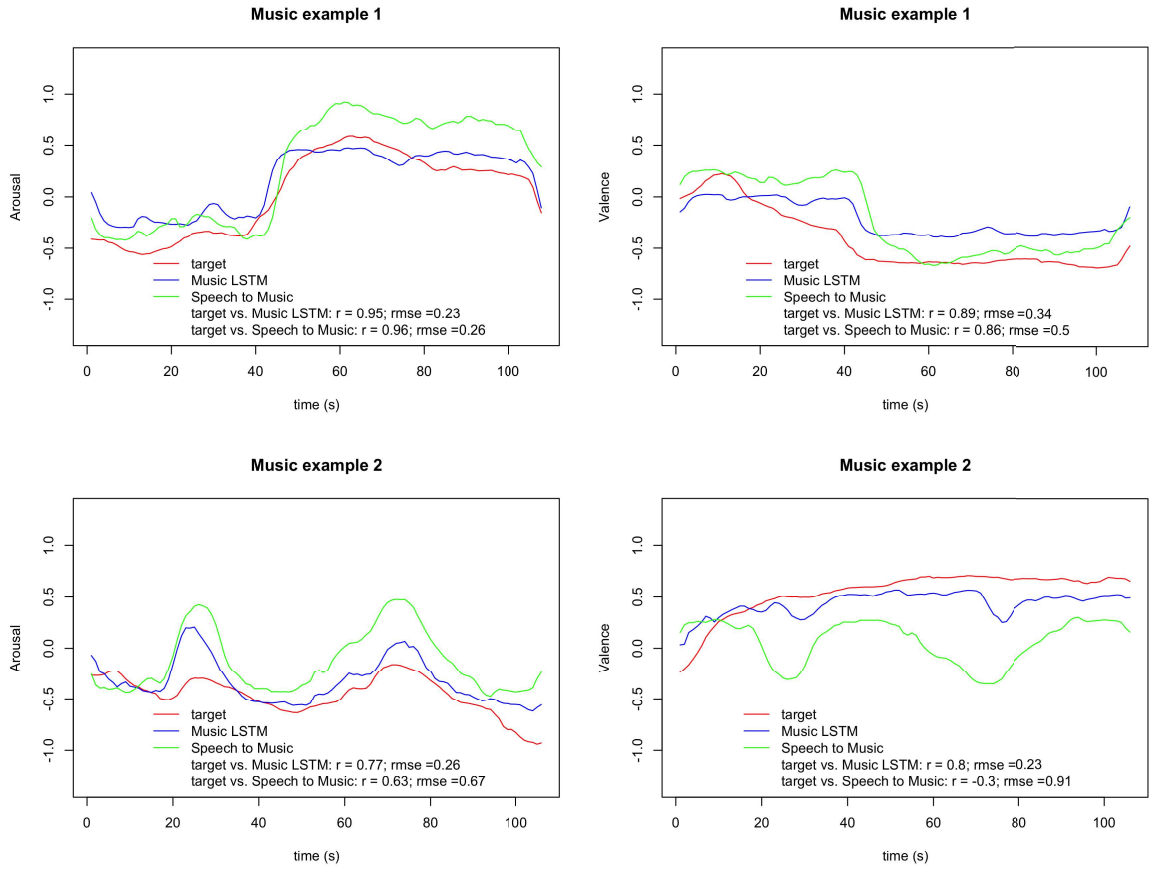

Fig. 2. Results: comparison between target data and models' predictions of Arousal and Valence for two representative instances of the Music data set.

[2] A. Gabrielsson and E. Lindström, "The role of structure in the musical expression of emotions," in *Handbook of music and emotion: Theory, research, applications*, P. N. Juslin and J. Sloboda, Eds. Oxford: Oxford University Press, 2010, pp. 367–400.

[3] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: different channels, same code?" *Psychological bulletin*, vol. 129, no. 5, pp. 770–814, Sep. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12956543

[4] G. Ilie and W. F. Thompson, "Experiential and Cognitive Changes Following Seven Minutes Exposure to Music and Speech," *Music Perception: An Interdisciplinary Journal*, vol. 28, no. 3, pp. 247–264, Feb. 2011. [Online]. Available: http://www.jstor.org/stable/info/10.1525/mp.2011.28.3.247

[5] E. Coutinho and N. Dibben, "Psychoacoustic cues to emotion in speech prosody and music." *Cognition & emotion*, pp. 1–27, Oct. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23057507

[6] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common," *Frontiers in Psychology*, vol. 4, no. May, pp. 1–12, 2013.

[7] E. Coutinho, "Computational and Psycho-Physiological Investigations of Musical Emotions," Ph.D. dissertation, University of Plymouth (UK), 2008.

[8] E. Coutinho and A. Cangelosi, "The Use of Spatio-Temporal Connectionist Models in Psychological Studies of Musical Emotions," *Music Perception*, vol. 27, no. 1, pp. 1–15, Sep. 2009. [Online]. Available: http://caliber.ucpress.net/doi/abs/10.1525/mp.2009.27.1.1 http://www.jstor.org/stable/40286139

[9] E. Coutinho and N. Dibben, "Music, Speech and Emotion: psychophysiological and computational investigations," in *Proceedings of the International Conference on Interdisciplinary Musicology: Nature versus Culture (CIM'10)*, R. Timmers and N. Dibben, Eds. Sheffield: University of Sheffield, 2010, pp. 47–48.

[10] E. Coutinho and A. Cangelosi, "Musical Emotions : Predicting Second-by-Second Subjective Feelings of Emotion From Low-Level Psychoacoustic Features and Physiological Measurements," *Emotion*, vol. 11, no. 4, pp. 921–937, 2011.

[11] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies." in *INTERSPEECH*, 2008, pp. 597–600.

[12] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," in *Affective computing and intelligent interaction*. Springer, 2007, pp. 488–500.

[13] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 867–881, 2010.

[14] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456.

[15] J. Chalupper and H. Fastl, "Dynamic loudness model (dlm) for normal and hearing-impaired listeners," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 378–386, 2002.

[16] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically." *Behavior research*

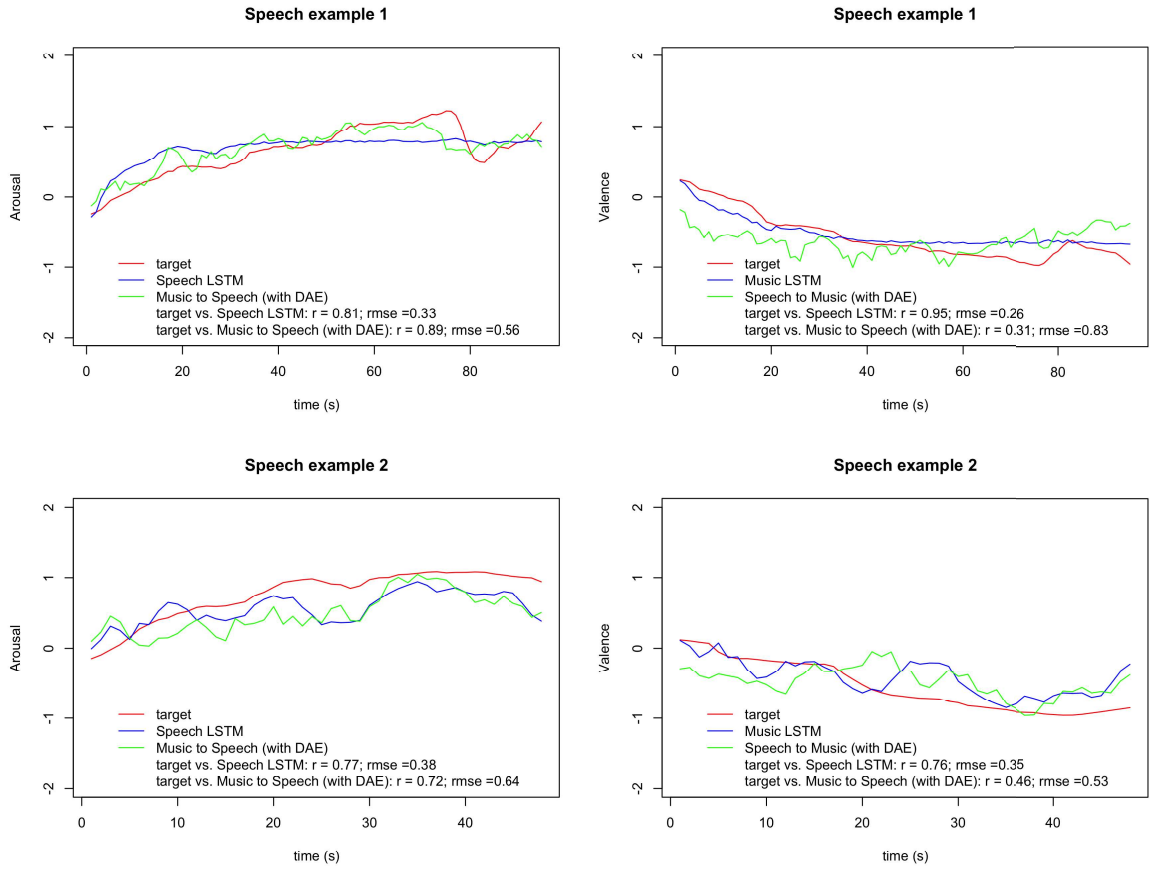

Fig. 3. Results: comparison between target data and models' predictions of Arousal and Valence for two representative instances of the Speech data set.

methods, vol. 41, no. 2, pp. 385–90, May 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19363178

[17] S. Dixon, "Evaluation of the Audio Beat Tracking System BeatRoot," *Journal of New Music Research*, vol. 36, no. 1, pp. 39–50, Mar. 2007. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/09298210701653310

[18] P. Mertens, "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Speech Prosody 2004, International Conference*, 2004.

[19] C. Dittmar, K. Dressler, and K. Rosenbauer, "A toolbox for automatic transcription of polyphonic music," in *Proceedings of audio mostly: 2nd conference on interaction with sound*. Citeseer, 2007, pp. 58–65.

[20] W. Aures, "Ein berechnungsverfahren der rauhigkeit," *Acta Acustica united with Acustica*, vol. 58, no. 5, pp. 268–281, 1985.

[21] P. N. Vassilakis, "Perceptual and physical properties of amplitude fluctuation and their musical significance," Ph.D. dissertation, UNIVERSITY OF CALIFORNIA Los Angeles, 2001.

[22] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980. [Online]. Available: http://content.apa.org/journals/psp/39/6/1161 http://psycnet.apa.org/journals/psp/39/6/1161

[23] E. Schubert, "Modeling perceived emotion with continuous musical features," *Music perception*, vol. 21, no. 4, pp. 561–585, 2004.

[24] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[25] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[26] A. Graves, *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, vol. 385.

[27] M. Wöllmer, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Tandem decoding of children's speech for keyword detection in a child-robot interaction scenario," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, p. 12, 2011.

[28] M. Wollmer, C. Blaschke, T. Schindl, B. Schuller, B. Farber, S. Mayer, and B. Trefflich, "Online driver distraction detection using long short-term memory," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 2, pp. 574–582, 2011.

[29] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *Affective Computing, IEEE Transactions on*, vol. 1, no. 2, pp. 119–131, 2010.

[30] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. ACII* , HUMAINE Association. Geneva, Switzerland: IEEE, 2013, pp. 511–516.

[31] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," 2013.

[32] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[33] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 513–520.