Predicting Emotions Induced by Music Using System Identification Theory

Mahdi Khajehim
Faculty of Biomedical Engineering
Amirkabir University of Technology
Tehran, Iran
mahdi.khajehim@aut.ac.ir

Sahar Moghimi
Department of Biomedical Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
s.moghimi@um.ac.ir

Abstract—**Modeling the emotional content of music is of great importance, since it is believed that music is capable of inducing different emotions. In this study we present an Autoregressive with Exogenous Input (ARX) model based on system identification theory for modeling the emotional content of music in a two dimensional emotion space and also a nonlinear Autoregressive with Exogenous Input (NARX) model to capture the nonlinear characteristics of the system. We also investigate the causal relationship between musical features and the induced emotions by removing the autoregressive terms from the developed model. Finally A brief discussion about the most important features is presented.**

Keywords- emotional content; system identification; music; valence; intensity

I. INTRODUCTION

Nowadays With rapid evolution of new technologies, perception of human emotions is a key factor in human-computer interaction (HCI). Machines must be able to evaluate user's emotion (e.g. emotions about a movie, music, or a game) in an easy and reliable way. This is necessary for a better mutual sympathy between machines and humans. Perception of the emotional content of music is an interesting and relatively fresh subject of study in psychology and music analysis. Obviously it is a complex issue influenced by culture, gender, age and many other parameters. However, if we consider music as a carrier for emotions, it makes more sense to focus on the music itself rather than how an individual will interpret its emotional content. So far several researches have been carried out in this field. Feng et al. [1] Li and Ogihara[2], and Vaizman et al.[3] used some musical features to extract different emotions, but they assumed emotions as a discrete variable. Since music is continuous and time-varying in nature, by expressing music with a discrete emotion we unreasonably eliminate the continuous properties of the system. Schubert suggested considering the emotions as a continuous variable [4]. He presented a mathematical model for musical emotions based on music features. Korhonen [5] also used this method to construct a generalized model for emotional content of music as a continuous variable. Schmidt and Kim [6] developed a second by second emotional labeling method in a two dimensional emotion space as another solution. Here we use a system identification approach for modeling this emotional content in a two dimensional emotion space as a continuous variable. We first introduce a better Autoregressive with Exogenous Input (ARX) model in comparison to previous researches, and then proceed by developing a novel nonlinear Autoregressive with Exogenous Input (NARX) model with improved results. We discuss the prominent drawbacks of these models and present a solution for it. Finally a discussion about the most important features, based on a sequential forward selection algorithm is presented.

II. BACKGROUND INFORMATION

A. Measuring emotions

Generally there are two major ways for evaluating emotions. One approach is to choose emotions from a predefined list of words (e.g. happiness, sadness, fear). Many researches aimed at recognizing these emotions using physiological signals [7, 8] , however this method is applicable only if we consider emotions as a discrete variable. There is not a unanimous agreement about the exact meaning of emotions and the best words to describe them [9]. The second approach is to measure emotions in a two dimensional emotion space, suggested by Lang et al. [10] and in a similar way by Russell [11]. The two dimensions are intensity and valence; with intensity representing the intensity of the experienced emotion, and valence describing the degree of the pleasantness (Fig. 1). Emotions can be described here by valence-intensity values: e.g. sadness has negative valence and low intensity or happiness has positive valence and high intensity. Now by continuous recording of these two-axis values, we are capable of measuring emotions induced by external stimuli over time. Since in this research we assumed emotions as a continuous variable of time, second approach was chosen.

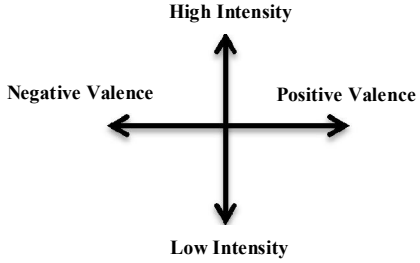

Figure 1. Two-dimensional model for emotion

978-1-4799-3232-0/13/$31.00 ©2013 IEEE

Table I. MUSICAL SELECTIONS USED IN THIS PAPER

Musical Selection	Composer
Concertio de Aranjuez(Adagio)	Rodrigo
Morning(Peer Gynt)	Grieg
Piano Concerto NO.1 (Allegro Maestoso)	Liszt
Pizzicato Polka	J.Strauss

B. Database information

Here we use the database collected by Korhonen et al. [12]. It contained 6 musical pieces. In this work we discarded two musical pieces of this collection due to their somehow neutral emotional content, which in our experiments as well as Korhonen's study [5] resulted in misleading predictions. Table I shows the musical selections we used in this study. Data are collected from 35 persons and each person listened to all musical selections in a random order. For measuring induced emotions two programs called FEELRTACE [13] and EmotionSpace Lab [4] were used. The user can describe emotions by using a pointer device in these programs. Measured emotions are recorded as normalized valence-intensity values over time. Input data are 18 musical features collected by two programs, PsySound [14] and Musical Research System for Analysis and Synthesis (MARSYAS) [15] based on musical properties including Dynamics, Timber, Harmony and Texture.

III. PROPOSED MODELS

A. ARX Model

In order to improve the results reported by Korhonen et al. [12], we tried to develop a more efficient ARX model based on the classical Gram-Schmidt orthogonalization algorithm [16].

From system identification theory [17], a generalize ARX model for a multi-input single-output (MISO) system is:

$$\begin{aligned} y(t) =& a_1 y(t-1)+a_2 y(t-2)+\ldots+a_n y(t-n) \\ &+b_{11}u_1(t-1)+\ldots+b_{1m}u_1(t-m) \\ &\vdots \qquad\qquad \vdots \\ &+b_{k1}u_k(t-1)+\ldots+b_{km}u_k(t-m)+e(t)\ , \quad \text{for } t=1,\ldots,N \end{aligned} \tag{1}$$

Where y(t) is the output, n and m are the number of output and input regressors, respectively , k shows the number of inputs , N is the data length and e is the prediction error. If we consider θ as follows [16]:

$$\theta = [a_1\ a_2\ \ldots\ a_n\ b_{11}\ \ldots\ b_{1m}\ \ldots b_{k1}\ \ldots b_{km}]^T \tag{2}$$

and also:

$$X=[y(t-1)\ \ldots\ y(t-n)\ \ u_1(t-1)\ \ldots\ u_k(t-m)]^T \tag{3}$$

we have:

$$y(t)=X^T(t)\theta \tag{4}$$

If we consider M as the number of parameters, then assuming:

$$X=W.A \tag{5}$$

$$W=[w_1\ w_2\ \ldots w_M] \tag{6}$$

$$A=\begin{bmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots & \vdots & a_{2M} \\ 0 & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & a_{M-1M} \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \tag{7}$$

Where A is an upper M×M triangular matrix and W is a matrix with orthogonal columns. W is built with the following formulation and through the orthogonal Gram-Schmidt decomposition. At the kth step of the algorithm:

$$w_1=x_1 \tag{8}$$

And for k=2,…, M:

$$a_{ik}=\frac{<w_i,x_k>}{<w_i,w_i>} \quad , \quad 1 \le i<k \tag{9}$$

$$w_k=x_k-\sum_{i=1}^{k-1} a_{ik}w_i \tag{10}$$

where <w,x> denotes inner product. Now we define:

$$D=W^TW \tag{11}$$

$$g=D^{-1}W^Ty \tag{12}$$

Finally, at the end of the evaluation phase we can calculate θ from the training dataset as follows:

$$\theta=A^{-1}g \tag{13}$$

And the estimated output (Z) is as follows:

$$Z=X\theta+e \tag{14}$$

More comprehensive details and formula for calculating fitting can be found in [16, 17].

Two different ARX models for valence and intensity were developed. Table II shows the fitting values obtained from the validation phase.

An improvement was observed in the prediction of valence from what was previously reported. Korhonen achieved an average fitting of 21% which in this study reached to 48.2%. Lower fitting values were obtained for intensity Korhonen achieved 78% average fitting, while ours was 67%. However, by taking into account both the valence and intensity values a better overall fitting was achieved compared to what was reported by Korhonen.

B. NARX Model

Since emotion perception for an individual depends on many parameters, it seems that a nonlinear approach can be a better choice for modeling this influence of music. Therefore we developed a NARX model in order to improve our previously introduced ARX model. Akaike Information Criterion (AIC) was used as a stopping criterion in the forward selection procedure. Utilizing AIC results in a Tradeoff between model complexity and flexibility.

The General form for a NARX model is (in a MISO system):

$$\begin{aligned} y(t) =& f(y(t-1),\ ,\ldots,y(t-n) \\ &u_1(t-1),\ldots,u_1(t-m) \\ &\vdots \qquad\qquad \vdots \\ &u_k(t-1),\ldots,u_k(t-m)) +\xi \quad \text{for } t=1,\ldots,N \end{aligned} \tag{15}$$

TABLE II. VALIDATION RESULTS FOR THE ARX MODEL

Musical selection	Valence Fitting (%)	Intensity Fitting (%)
Morning (Peer Gynt)	63	80
Allegro Maestoso (Piano Concerto No.1)	71	66
Pizzicato Polka	42	50
Concierto de Aranjuez (Adagio)	17	72

where $f(.)$ is a nonlinear function, n and m are the number of output and input regressors, k is the number of inputs and ξ is the modeling error. Since any $f(.)$ can be approximated with a polynomial, so we have:

$$y(t) = \sum_{i=1}^{M} p(t)\,\theta_i + \xi(t) \quad ,t = 1,\ldots,N \tag{16}$$

where y(t) is the output, p_i (t) is the monomials of X columns in (3), θ contains the estimated parameters. This equation can be written in the matrix form [16]:

$$Y = P\Theta + \Xi \tag{17}$$

where:

$$Y=\begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix},\ P=[p_1 \quad \ldots \quad p_M],\ \Theta=\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix},\ \Xi=\begin{bmatrix} \xi(1) \\ \vdots \\ \xi(M) \end{bmatrix} \tag{18}$$

and M is the number of parameters, and also:

$$p_i=\begin{bmatrix} p_{i(1)} \\ \vdots \\ p_{i(N)} \end{bmatrix} \tag{19}$$

Since only a relatively small number of regressors suffice for adequate characterization of the system dynamics [16]. $\overline{\Theta}$ can be calculated by minimizing $|Z-P\Theta|$ in a least square framework.

We use the orthogonal Gram-Schmidt with forward selection to find the most effective regressors.by considering:

$$P = WA \tag{20}$$

Where W and A are similar to what was described in the ARX model. For each regressor at the kth step(k=2,…,n_s) and for i=1,…,n with $i \neq i_1,\ldots,i_{k-1}$ [18]:

$$a_{jk}^{(i)} = \frac{w_j^T p_i}{w_j^T w_j} \ , \text{ with } j=1,\ldots,k-1 \tag{21}$$

$$w_k^{(i)} = p_i - \sum_{j=1}^{k-1} a_{jk}^{(i)}\, w_j \tag{22}$$

$$g_k^{(i)} = \frac{w_k^{(i)T} y}{w_k^{(i)T} w_k^{(i)}} \tag{23}$$

$$err_k^{(i)} = \frac{(g_k^{(i)})^2 w_k^{(i)T} w_k^{(i)}}{y^T y} \tag{24}$$

We found and utilized the regressor associated with the largest error reduction ratio (err).

An AIC was used as the stopping criterion during the forward subset selection to terminate forward subset selection procedure [18].

Table III shows the results obtained from the validation phase using the developed NARX model. This table illustrates better average value for intensity average fitting (78%) compared to the ARX model, and almost equal results for valence predication(fitting of 46%). It was observed that In the forward selection procedure in most cases y(t-1) was selected as the most important regressor. In the next step we aimed at Investigating the causal relationship between the musical features and the self-evaluated emotions. This would allow us to decide whether we can predict the induced emotion solely based on stimulus characteristics. Moreover one cannot expect autoregressive term to be available in practical applications. Here, another model was developed by omitting the output past terms from the regressor pool. All the theoretical details in this case are similar to what was previously stated for NARX model .The modified model was not able to predict the valence values with acceptable accuracy.

TABLE III. VALIDATION RESULTS FOR THE NARX MODEL

Musical selection	Valence Fitting (%)	Intensity Fitting (%)
Morning (Peer Gynt)	67	82
Allegro Maestoso (Piano Concerto No.1)	78	83
Pizzicato Polka	57	68
Concierto de Aranjuez (Adagio)	-19	79

Table IV illustrates the results obtained for predicting the intensity values during the validation phase. Fig. 2 shows an example of estimated and actual values of the reported induced intensity by Morning (Peer Gynt).One of the reasons for the poor performance of the model in predicting valence can be explained as follows: we believe that the intersubject variation of the reported valence was larger than that of intensity, since valence can be modulated by factors such as extra association, mood, and amount of exposure to a specific music genre in the subject's daily life, while it is widely reported in the literature that intensity can be modulated by the violation of expectancy.

A closer look at the selected regressors in the identification phase revealed that the Short Term Max Loudness (STML) [14], which is a dynamic feature, was selected in almost all cases as the most important regressor. This demonstrates that STML and its related regressors play an important role in estimating the values of the induced intensity.

IV. CONCLUSION AND SUGGESTION

In this paper we investigated the emotional content of music and the application of two parametric models for predicting this content. A linear as well as a nonlinear ARX model was practiced for prediction of the reported valence and intensity. The proposed NARX model was further modified in order to explore the role of the utilized musical features for predicting the reported intensity. The proposed model was able to follow the pattern of the reported intensity solely based on the musical features (Fig .2).

For future researches a more comprehensive look at the effective inputs for predicting intensity and valence is required. Although the proposed model was not able to determine the causal relationship between the utilized musical features and valence, a more through experiment with a wider musical selection from different genres may lead to a better performance. Since valence is highly dependent on the personal comprehension and judgment of the stimuli, electrophysiological as well as peripheral signals, recorded from the subject during the test procedure, may be better for predicting the experienced valence.

TABLE IV.RESULTS FROM SIMULATION MODEL

Musical selection	Intensity Fitting (%)
Morning (Peer Gynt)	42
Allegro Maestoso (Piano Concerto No.1)	43
Pizzicato Polka	-64
Concierto de Aranjuez (Adagio)	42

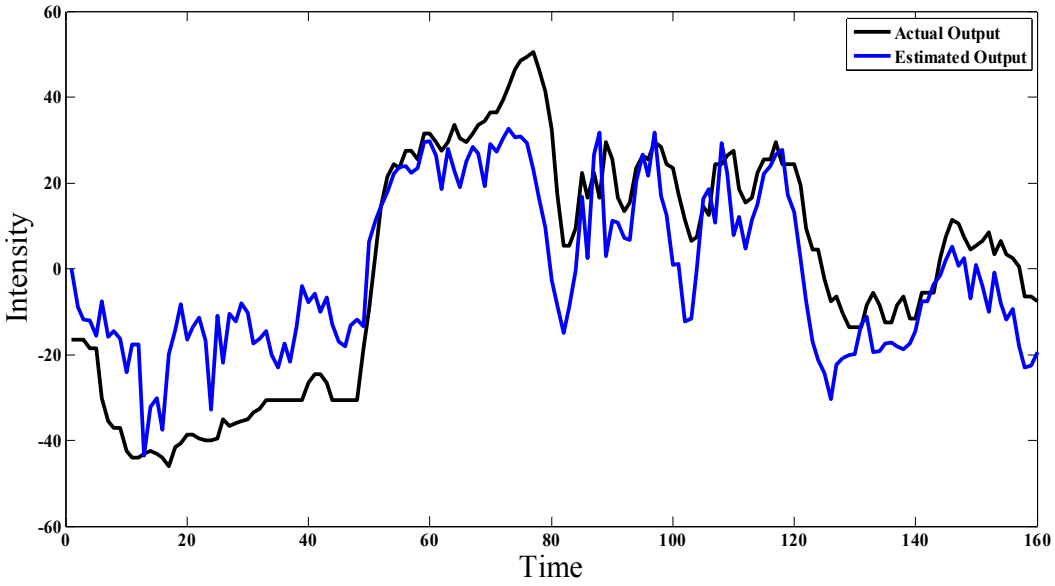

Figure 2. Estimated and actual output for causal model (intensity of Morning)

REFERENCES

[1] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," *In Proc. 26th Annu. Int. ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR)*, Toronto, ON, Canada, pp.375-376, Jul.2003.

[2] T. Li and M. Ogihara, "Detecting Emotion in Music "Detecting emotion in music." In *ISMIR*, vol. 3, pp. 239-240, 2003.

[3] Y. Vaizman, R. Y. Granot, G. Lanckriet, " Modeling Dynamic Patterns for Emotional Content in Music," In *ISMIR*, pp. 747-752. 2011.

[4] E. Schubert, "Measuring emotion continuously: Validity and reliability of the two- dimensional emotion space," *Aust. J. Psychol*, vol.51, no.3, pp.154-165, Dec. 1999.

[5] M. D. Korhonen, "Modeling continuous emotional appraisals of music using system identification," M.S. thesis, Syst. Des. Eng. Univ. Waterloo, ON, Canada, 2004.

[6] E.M. Schmidt, Y. E. Kim, "Prediction of Time-Varying Musical Mood Distributions from Audio," In *ISMIR*, pp. 465-470, 2010.

[7] J. Kim, E. Andre, "Emotion recognition based on Physiological changes in Music Listening," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no .12, pp. 2067-2073, Dec.2008.

[8] Y. P. Lin, C.H. Wang, T. L. Wu, S. K. Jeng, J. H. Chen, "EEG-based emotion recognition in music listening: a comparison of schemes for multiclass support vector machine," *Int Conf. on Acoustics, Speech, and Signal Processing*, pp.489-492, 2009.

[9] R. Cowie et al. "Emotion Recognition in Human-Computer interaction," *IEEE Signal Processing Mag*, pp. 32-80, Jun. 2001.

[10] P. J. Lang, "The Emotion Probe: Studies of Motivation and Attention," *American Psychologist*, vol. 50, pp. 372-385, 1995.

[11] J. A. Russell, "Measures of emotion," *In Emotion: Theory Research and Experience*, vol. 4, R. Plutchik and H. Kellerman, Eds. New York: Academic, 1989, pp. 81–111.

[12] M. D. Korhonen, D. A. Clausi, M. E. Jernigan, "Modeling Emotional Content of Music Using System Identification," *IEEE Trans. System Man and Cybernetics*, vol.36, no. 3 , pp. 588-599, Jun.2006.

[13] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELTRACE': An instrument for recording perceived emotion in real time," *in Proc. Speech and Emotion, ISCA Tutorial and Research Workshop (ITRW)*, Newcastle, U.K., Sep. 2000, pp. 19–24.

[14] Cabrera, Densil, "PSYSOUND: A computer program for psychoacoustical analysis." In *Proceedings of the Australian Acoustical Society Conference*, vol. 24, pp. 47-54, 1999.

[15] G. Tzanetakis and P. Cook, "MARSYAS: A framework for audio analysis," *Organ. Sound*, vol. 4, no. 3, pp. 169–175, 2000.

[16] S. Chen, S. A. Billings, W. Luo, "orthogonal least square methods and their application to nonlinear system identification," *Int. J. Control*, vol. 50, no. 5, pp.1873-1896, 1989.

[17] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[18] O. Nelles, *Nonlinear System Identification: from Classical Approaches to Neural Networks and Fuzzy Models*, springer, 2001.

978-1-4799-3232-0/13/$31.00 ©2013 IEEE