INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speech Based Emotion Prediction: Can a Linear Model Work?

Anda Ouyang, Ting Dang, Vidhyasaharan Sethu, Eliathamby Ambikairajah

School of Electrical Engineering and Telecommunications, UNSW, Sydney, Australia
{a.ouyang, ting.dang, v.sethu, e.ambikairajah}@unsw.edu.au

Abstract

Speech based continuous emotion prediction systems have predominantly been based on complex non-linear back-ends, with an increasing attention on long-short term memory recurrent neural networks. While this has led to accurate predictions, complex models may suffer from issues with interpretability, model selection and overfitting. In this paper, we demonstrate that a linear model can capture most of the relationship between speech features and emotion labels in the continuous arousal-valence space. Specifically, an autoregressive exogenous model (ARX) is shown to be an effective backend. This approach is validated on three commonly used databases, namely RECOLA, SEWA and USC CreativeIT, and shown to be comparable in terms of performance to state-of-the-art LSTM systems. More importantly, this approach allows for the use of well-established linear system theory to aid with model interpretability.

Index Terms: speech-based emotion prediction, autoregressive model, ARX models, affective computing

1. Introduction

Speech emotion recognition plays a vital role in a wide range of applications, such as in the development of intelligent devices, or in the clinical usages serving as mental state assessment tools [1] and has become a popular research area in recent years. Generally, an emotion prediction system consists of a frontend which extracts the emotion-related content from speech referred as features, and a backend that captures the relationship between features and emotion labels. Over the years, a large number of frontends have been developed for emotion prediction systems [2], with the acoustic feature set of eGeMAPS [3] being one of the more commonly adopted ones. The nature of the backend is dictated to a large extent by the nature of the emotion labels, which typically may be one of three major representations of human emotion states advocated by psychologists. Namely, categorical, dimensional and appraisal based representations [4]. In this work we focus on the dimensional representation of emotion which tends to better reflect the temporal dynamics in affective interaction [5]. It describes emotion by a bipolar circumplex model consisting of arousal (degree of activation level) and valence (degree of pleasant level) in terms of numerical values and leads to the use of suitable regression models as the backend.

While a significant number of regression modelling techniques have been proposed for emotion predictions, the most widely adopted models are Support Vector Regression [6] owing to its robustness, and Long-short term memory recurrent neural network (LSTM-RNN) [7-9] which captures the temporal dynamics of emotion. In addition, relevance vector machines (RVM) [10] and Gaussian mixture regression (GMR) [11] have also shown good performance in emotion prediction tasks. However, these models: 1) are either complex nonlinear models which involves a large number of model parameters that need to be estimated, such as LSTM-RNN and GMR; 2) suffer from significant increasing computation time with a relatively small increase on training data size such as SVR; 3) or ignores the temporal dependencies within emotion labels such as with most RVM and SVR based systems. In addition, the lack of interpretability of these non-linear models also limited the development of emotion prediction systems.

The use of linear models in the backend would help with interpretability but current wisdom suggests that they may not be adequate to model the complex relationship between speech and emotions. This is supported by the observation that previous studies on linear models in speech emotion recognition such as multivariate linear regression [12] only focus on the emotion content at current time step; and linear model that take into account temporal dynamics such as autoregressive exogenous (ARX) models [3] have only been invested in the context of categorial emotion classification systems [13]. In this paper we demonstrate that ARX models can serve as suitable backends for the continuous prediction of emotion labels in the arousal-valence space and additionally lend themselves to a variety of analyses that aid with interpretability of the model.

2. Proposed Linear Approach

In this section, we describe how the continuous emotion prediction task can be formulated as a linear regression system using an ARX model, where the current emotional state (either arousal or valence) can be predicted as a combination of the past sequence of arousal (or valence) states and the input speech features.

2.1. Model description

The ARX model formulates the arousal (or valence) value at time t as a linear combination of past arousal (or valence) values as well as current and past input speech features as follows:

$$y(t) = \sum_{i=1}^{n_a} a_i y(t-i) + \sum_{j=n_d}^{n_b} \boldsymbol{b}_j^T \boldsymbol{u}(t-j) \qquad (1)$$

where $y(t)$ denotes the arousal (or valence) at time frame t; $\boldsymbol{u}(t)$ denotes input speech feature vector corresponding to frame t; $\boldsymbol{b}_j$ denotes the vector of weights corresponding to each feature frame j, consisting of weight scalars for each feature dimension; n_a represents the order of AR sub-model; n_d is the delay between input features and predicted output, typically referred to as evaluators' reaction lag [14]. It is generally compensated by manually shifting the feature frames backwards or the labels forward prior to model training, however the ARX model is able to incorporate the reaction lag

http://dx.doi.org/10.21437/Interspeech.2019-3149

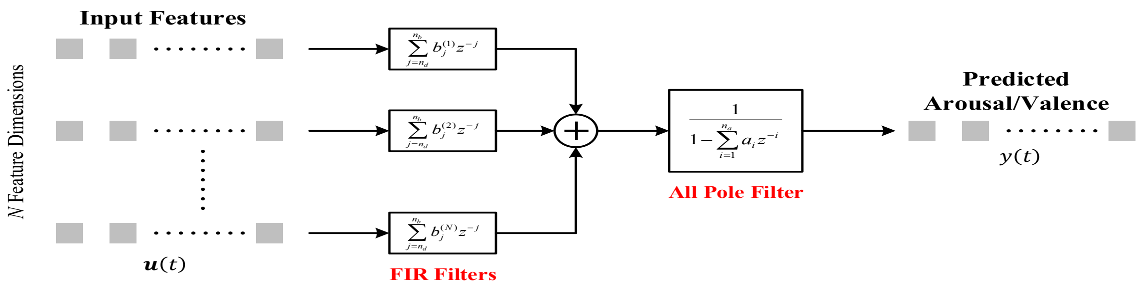

Input Features
N Feature Dimensions
$\sum_{j=n_d}^{n_b} b_j^{(1)} z^{-j}$
$\sum_{j=n_d}^{n_b} b_j^{(2)} z^{-j}$
$\sum_{j=n_d}^{n_b} b_j^{(N)} z^{-j}$
$\boldsymbol{u}(t)$
FIR Filters
$\frac{1}{1-\sum_{i=1}^{n_a} a_i z^{-i}}$
All Pole Filter
Predicted Arousal/Valence
$y(t)$

Figure 1: *Overview of proposed ARX back-end as a set of parallel FIR filters operating on each feature dimension in cascade with an all-pole filter. Note that the FIR filter coefficients are elements of the vector,* $\boldsymbol{b_j} = \left[b_j^{(1)}, b_j^{(2)}, \dots, b_j^{(N)}\right]^T$.

in the model directly. The order of the eXogenous sub-model is denoted by the number of time frames of considered past inputs which is equivalent to $n_b - n_d + 1$.

The ARX model can be regarded as a combination of a multivariate linear regression model (eXogenous) and an all-pole filter (AR model). In order to ascertain whether this model would be suitable and not result in spurious predictions we carried some preliminary statistical tests. Specifically, the Augmented Dickey-Fuller test [15] indicated that most of the annotation labels are most likely to be non-stationary, while all the speech features we employed (88-dimensional eGeMAPS) tended to be stationary. Furthermore, we also carried out some preliminary tests to confirm that spurious regression between non-stationary series, as suggested by Granger and Newbold [16], was not a significant factor in our system.

Finally, it should be noted that this model is essentially a set of parallel FIR filters (one for each dimension of the input feature dimension) connected in cascade with an all-pole filter, as shown in Figure 1. This enables the rich, well-established theory of linear systems to be brought to bear on the problem of model interpretability.

2.2. Parameter estimations

Given the linear prediction model outlined above, its parameters can be simply estimated as a least-squares estimate or a regularised least-squares estimate. In this work we adopt an L1 regularised least-squares estimate in order to encourage sparsity [17] in the parameters which can further help with model interpretability. Specifically, we estimate the parameters of the ARX model, $\widehat{\boldsymbol{\theta}}$, as:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{N_T} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \qquad (2)$$

where, λ is the regularization parameter which controls the strength of shrinkage with a larger λ encouraging greater sparsity; N_T denotes the length of the training data; $\boldsymbol{y}$ denotes a vector of all arousal (or valence) labels and can be written as $\boldsymbol{y} = \left[y_1, y_2, \dots y_{N_T}\right]^T$; $\boldsymbol{\theta}$ is the vector of ARX model parameters given as $\boldsymbol{\theta} = \left[a_1, a_2, \dots, a_{n_a}, \boldsymbol{b}_{\boldsymbol{n_d}}^{\boldsymbol{T}}, \boldsymbol{b}_{\boldsymbol{n_d+1}}^{\boldsymbol{T}}, \dots, \boldsymbol{b}_{\boldsymbol{n_b}}^{\boldsymbol{T}}\right]^T$; and $\mathbf{X}$ is a matrix of training data with the t^{th} row of $\mathbf{X}$ given by:

$$\boldsymbol{x_t} = \left[\boldsymbol{y}_{\boldsymbol{t}}^{\boldsymbol{T}}, \boldsymbol{u}^{\boldsymbol{T}}(t-n_d), \dots \boldsymbol{u}^{\boldsymbol{T}}(t-n_b)\right]^T \qquad (3)$$

with $\boldsymbol{y_t} = [y(t-1), \dots y(t-n_a)]^T$ denoting the sub-vector of $\boldsymbol{y}$ comprising of the n_a elements prior to frame t.

The recursive Alternating Direction Method of Multipliers (ADMM) [18] was chosen to learn the best-fit parameter array, owing to its computational efficiency and adaptability to convex optimizations like Lasso regression [18].

3. Database and Experimental Settings

3.1. Databases

Three databases were used in the experiments reported in this paper, namely, the RECOLA [19], the USC CreativeIT [20] and the SEWA [21] databases. The RECOLA database and the SEWA database consist of spontaneous data, and a selected subset of these two databases are used as per Audio/Vision Emotion Challenge (AVEC) 2016 [22] and AVEC 2017 [21]. 18 utterances with each of 5 minutes in the RECOLA database and 48 utterances in the SEWA database are adopted in this paper. The USC CreativeIT database contains acted data, where 8 sessions including 90 utterances are used.

The RECOLA and the SEWA databases were divided into training and development sets as per the AVEC 2016 [22] and AVEC 2017 [21] respectively, while USC CreativeIT database was organized in a leave-one-session-out cross validation form, as per [23].

3.2. Experimental setup

The eGeMAPS features are employed as the input speech features and were extracted using OpenSmile [24]. The features were extracted every 40ms and 100ms to match the annotation rates in the RECOLA and the SEWA databases respectively. The USC CreativeIT database employs an annotation rate of 60hz and eGeMAPS features cannot be directly extracted at that rate using OpenSmile. Consequently, the eGeMAPS features for this database were instead extracted every 10 milliseconds (100Hz), and 6 out of 10 frames were then taken to match the 60 Hz annotation rate, as per [25]. Finally, a moving average filter is applied to both the features and annotations to down sample both from 60Hz to every 1Hz as in [11, 26]. This is accordance with the suggestion that emotion prediction at a larger window level is less noisy and more robust compared to smaller frame-level prediction [27]. Finally, feature normalization was applied to each feature dimension. This is carried out (for both training and test datasets) by subtracting the mean and dividing the standard deviation, which are both estimated from the training set.

For model parameter estimation, we use Alternating Direction Method of Multipliers algorithm (ADMM) [18] implemented in the Lasso Toolbox in MATLAB. The parameters of the ARX model and regularization parameter, λ, are optimized independently. Firstly, a grid search was employed (with λ set to 0) in order to find suitable model hyperparameters, $\{n_a, n_d, n_b\}$, based on prediction accuracy on the development sets. Following this, a second grid search was employed to ascertain a suitable value for λ.

During the test phase, the output predictions from the ARX model are normalised based on the first two moments of the training labels. Following this the accuracy of the model is quantified using concordance correlation coefficient (CCC) as the performance metric for both arousal and valence. It should be noted that one advantage of ARX model is that autoregressive filter in ARX model, which regresses to a non-trivial low pass IIR filter as described in section 4.3 and acts as a smoothing filter, optimised for the task at hand, that eliminates the need for additional post processing.

4. Preliminary Analysis

In this section, a variety of analyses carried out to explore the interpretability of the model are reported. We quantify the temporal dependencies of the predicted emotional state on input feature and past emotional states by examining the estimated AR model and X model orders respectively. Further, the feature dimensions that are consistently chosen as significant components for arousal and valence predictions are identified. Finally, since the ARX model can be decomposed into an AR filter operating in the space of emotion labels and a set of parallel FIR filters, each operating on a distinct feature dimension, some preliminary analyses of their frequency responses are also carried out.

4.1. Temporal dependencies

Figure 2 depicts the CCC obtained when predicting valence on the RECOLA database with varying eXogenous (X) and AR model orders, which in turn correspond to the durations of inputs and past predictions respectively on which the current prediction is based. In Figure 2, AR model order starts from 0 and the X order starts from 1, which corresponds to a pure multivariate linear regression system with a fixed delay between input and output. As the AR model order is then increased from 0 to 7, it can be observed that there is a significant improvement in the model with CCC increasing from 0.189 to 0.421. Further increase in model accuracy with the increase of X model order from 0 to 15 is also evident. These findings imply that valence intensity is highly dependent on the past valence intensities, as well as the past speech characteristics. Finally, it should be noted that the performance surface is smooth and fairly flat which suggests that model is not overfitting and can be expected to work over a fairly broad range of AR and X model order choices. Similar trends were also observed for both arousal and valence prediction systems on all three databases.

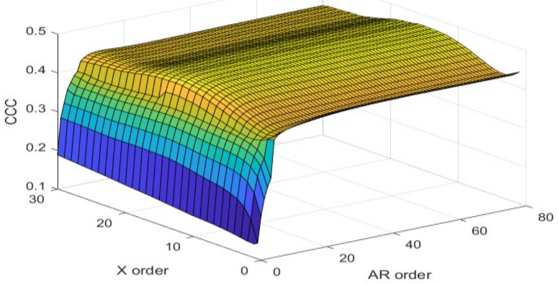


Figure 2: *Grid search in RECOLA's valence*

The observation that the accuracy of the model remains fairly constant for AR model orders greater than 10, while a small but distinct optima can be observed at a X model order of 15 (refer Figure 2) suggesting that X order is more sensitive than AR order when selecting the most appropriate model parameters. Therefore, in subsequent experiments we first choose the X order which maximizes CCC and then the minimum AR order that still achieves a good CCC, aiming to minimize the model complexity. The selected optimal parameters in three databases are listed in Table 1. In this table A and V denote arousal and valence respectively and the reported delay corresponds to n_d in equation (1).

Table 1: *AR and X orders among three databases*

Database	**RECOLA**		**CreativeIT**		**SEWA**	
	A	V	A	V	A	V
AR order	13	15	2	2	11	11
X order	8	15	1	2	2	2
Delay (n_d)	50	7	0	0	0	0

It is worth noting that the time interval between predictions typically employed on all three databases are different. This in turn means that for the RECOLA database, AR orders of 13 and 15 for arousal and valence corresponds to 0.52 second and 0.6 second; for the SEWA database, AR orders for arousal and valence both correspond to 1.1 seconds in time; and for the CreativeIT, the AR orders correspond to a slight longer time of 2 seconds. In general, Table 1 indicates that the temporal dependencies in the label space are in the range of 0.5 to 2 seconds. Similarly, we can conclude the time dependencies in the feature space are in the range of 0.2 to 2 seconds, matching that in the label space. Having said that, Figure 2 suggests that the model is not too sensitive to changes in the model order over a fairly broad range.

4.2. Feature Analysis

As discussed in section 2.2, L1 regularization is applied during parameter estimation, with the regularization term λ controlling the sparsity of the estimated weights. This in turn lets us increase λ and consequently decrease the number of features selected (when all associated weights go to zero) keeping only the most significant ones. This procedure lets us estimate which features contribute the most to emotion prediction. As the eGeMAPS feature set is composed of functionals, the exact feature dimensions may vary from database to database, but similar feature types were still observed in all three cases. For instance, pitch and MFCC related features were selected for both arousal and valence predictions on all three databases; and mean spectral flux features are selected with the highest weights for all arousal prediction systems across three databases and valence prediction systems on the RECOLA and SEWA databases. This suggests that these speech characteristics are the most emotion-specific features.

4.3. Frequency Spectrum

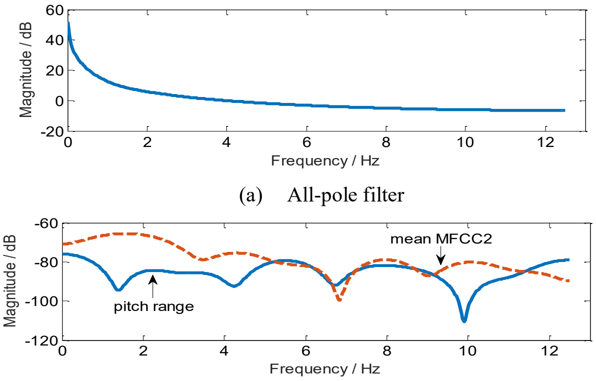
The structure of the ARX model (refer Figure 1) also gives us the opportunity to study the eXogenous and AR models in terms of the magnitude responses. Specifically, the AR components in equation (1) act as an all-pole filter with the transfer function $H_{AR}(z)$ given as:

$$H_{AR}(z) \; = \; \frac{1}{1-\sum_{i=1}^{n_a} a_i \, z^{-i}} \tag{4}$$

and the transfer functions of the FIR filter corresponding to n^{th} dimension of the input features, $H_n(z)$ is given as:

$$H_n(z) \; = \; \sum_{j=n_d}^{n_b} b_j^{(n)} \, z^{-j} \tag{5}$$

The magnitude response of all-pole filter is shown in Figure 3(a), and the magnitude response of the FIR filters for two important features, namely, the range between 20th and 80th percentile of pitch and mean MFCC2 are shown in Figure 3(b). It is seen from Figure 3(a) that the all-pole filter is a non-trivial low pass filter with a sharp transition band, which acts as a smoothing filter for predictions to some extent. The FIR filter responses corresponding to the two feature types differ from each other as can be seen from Figure 3(b), with the magnitude responses suggesting that the valence predictions are mainly related to the pitch information in frequency range of 0-0.5Hz, 5-6Hz and 7-9Hz, and to the MFCC2 variations in the low frequency range of 0-2.5Hz.



(a) All-pole filter

(b) FIR filters for pitch range and mean MFCC2

Figure 3: *Magnitude responses for valence prediction model trained on RECOLA.*

5. Experimental Results

The arousal and valence prediction accuracy of the proposed system employing a linear ARX model back-end in terms of CCC evaluated on all three databases are compared to the state-of-the-art non-linear models, including RVM, GMR and LSTM systems, and the results are shown in Tables 2-4 respectively.

Table 2: *System performance in the RECOLA database*

Features	Backend	CCC	
		Arousal	Valence
log Mel filterbank	LSTM [7]	0.859	0.596
LLD-based$_{(1)}$ functional based$_{(2)}$	LSTM [9]	0.785$_{(1)}$	0.378$_{(2)}$
log Mel filterbank	LSTM [8]	0.868	0.623
eGeMAPS	LSTM	0.599	0.363
eGeMAPS	ARX	0.783	0.467

* (1) and (2) indicate different feature sets used for arousal and valence predictions.

Table 3: *System performance in the SEWA database*

Features	Backend	CCC	
		Arousal	Valence
Soundnet$_{(1)}$ IS10$_{(2)}$	LSTM [28]	0.527$_{(1)}$	0.504$_{(2)}$
BottleNeck	LSTM [29]	0.533	0.466
eGeMAPS + BoAW	RVM [10]	0.494	0.507
eGeMAPS	ARX	0.540	0.502

Table 4: *System performance in the CreativeIT database*

Features	Backend	Arousal		Valence	
		CCC	CC	CCC	CC
LLDs	GMM [30]	-	0.565	-	0.507
LLDs	GMR [11]	0.405	0.617	0.155	0.266
eGeMA PS	ARX	0.444	0.469	0.215	0.240

It can be observed that ARX achieves comparable performances to the much more complex non-linear LSTM back-ends on all three databases. It is also worth noting that it outperforms the best LSTM system on the SEWA database, with 1.3% relative improvement for arousal. These results strongly suggest that a linear ARX model can capture a large fraction of the emotion specific information contained in the speech features and the non-linear models possibly only captures very limited additional information at the cost of significant increase in the model complexity.

6. Conclusion

This work set out to ask the question – "Can a linear model be used for speech based emotion prediction?". The results reported in this paper demonstrate that an emotion prediction system that employs a linear ARX model as its backend can exhibit good performance suggesting that the speech based emotion prediction task can indeed be approximated by a linear model. More importantly this work also demonstrates that by employing a linear model, a significantly richer suite of tools can be brought to bear to interpret the model. Some preliminary analyses based on well established linear system theory revealed that emotion denpendency for arousal and valence is in the range of 2 seconds, and the most significant and consistent features contributing to emotion predictions are those based on F0 and MFCC. Future work could be looking at adding non-stationary stochastic process models to capture information that may not be captured by linear models.

7. References

[1] B. Desplanques and K. Demuynck, "Cross-lingual Speech Emotion Recognition through Factor Analysis," in *Interspeech2018*, 2018: ISCA, pp. 3648-3652.

[2] P. Chandrasekar, S. Chapaneri, and D. Jayaswal, "Automatic Speech Emotion Recognition: A Survey," (in English), *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (Cscita),* pp. 341-346, 2014.

[3] L. Guo, D. Huang, and E. Hannan, "On ARX (∞) approximation," *Journal of Multivariate Analysis,* vol. 32, no. 1, pp. 17-47, 1990.

[4] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Face and Gesture 2011*, 2011: IEEE, pp. 827-834.

[5] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing,* vol. 3, no. 2, pp. 184-198, 2012.

[6] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017: IEEE, pp. 5005-5009.

[7] D. Le, Z. Aldeneh, and E. M. Provost, "Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent Neural Network," in *INTERSPEECH*, 2017, pp. 1108-1112.

[8] E. A. AlBadawy and Y. Kim, "Joint Discrete and Continuous Emotion Prediction Using Ensemble and End-to-End Approaches," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*, 2018: ACM, pp. 366-375.

[9] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017: IEEE, pp. 2367-2371.

[10] T. Dang *et al.*, "Investigating Word Affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017," presented at the Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, California, USA, 2017.

[11] T. Dang, V. Sethu, and E. Ambikairajah, "Compensation Techniques for Speaker Variability in Continuous Emotion Prediction," *IEEE Transactions on Affective Computing,* 2018.

[12] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "Speech Emotion Recognition: Methods and Cases Study," in *ICAART (2)*, 2018, pp. 175-182.

[13] D.-M. Yu and J.-A. Fang, "Research on a methodology to model speech emotion," in *2007 International Conference on Wavelet Analysis and Pattern Recognition*, 2007, vol. 2: IEEE, pp. 825-830.

[14] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing,* vol. 6, no. 2, pp. 97-108, 2014.

[15] R. I. Harris, "Testing for unit roots using the augmented Dickey-Fuller test: Some issues relating to the size, power and the lag structure of the test," *Economics letters,* vol. 38, no. 4, pp. 381-386, 1992.

[16] C. W. Granger, P. Newbold, and J. Econom, "Spurious regressions in econometrics," *A Companion to Theoretical Econometrics, Blackwell, Oxford,* pp. 557-561, 2001.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 58, no. 1, pp. 267-288, 1996.

[18] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning,* vol. 3, no. 1, pp. 1-122, 2011.

[19] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013: IEEE, pp. 1-8.

[20] A. Metallinou, Z. Yang, C.-c. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language resources and evaluation,* vol. 50, no. 3, pp. 497-521, 2016.

[21] F. Ringeval *et al.*, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017: ACM, pp. 3-9.

[22] F. Povolny *et al.*, "Multimodal emotion recognition for AVEC 2016 challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016: ACM, pp. 75-82.

[23] T. Dang, V. Sethu, and E. Ambikairajah, "Factor Analysis Based Speaker Normalisation for Continuous Emotion Prediction," in *INTERSPEECH*, 2016, pp. 913-917.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010: ACM, pp. 1459-1462.

[25] Z. Huang and J. Epps, "An Investigation of Partition-based and Phonetically-aware Acoustic Features for Continuous Emotion Prediction from Speech," *IEEE Transactions on Affective Computing,* 2018.

[26] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing,* vol. 31, no. 2, pp. 137-152, 2013.

[27] E. Bozkurt, Y. Yemez, and E. Erzin, "Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures," *Speech Communication,* vol. 85, pp. 29-42, 2016.

[28] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition," presented at the Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, California, USA, 2017.

[29] J. Huang *et al.*, "Continuous Multimodal Emotion Prediction Based on Long Short Term Memory Recurrent Neural Network," presented at the Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, California, USA, 2017.

[30] C.-M. Chang and C.-C. Lee, "Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion information," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017: IEEE, pp. 5820-5824.