Audio-Visual Emotion Forecasting: Characterizing and Predicting Future Emotion Using Deep Learning

Anonymous FG 2019 submission
Paper ID ****

***Abstract*— Emotion forecasting is the task of predicting the future emotion of a speaker—i.e., the emotion label of the future speaking turn–based on the speaker's past and current audio-visual cues. Emotion forecasting systems require new problem formulations that differ from traditional emotion recognition systems. In this paper, we first explore two types of *forecasting windows* (i.e., analysis windows for which the speaker's emotion is being forecasted): utterance forecasting and time forecasting. Utterance forecasting is based on speaking turns and forecasts what the speaker's emotion will be after one, two, or three speaking turns. Time forecasting forecasts what the speaker's emotion will be after a certain range of time, such as 1–5, 5–10, and 10–15 seconds. We then investigate the benefit of using the past audio-visual cues in addition to the current utterance. We design emotion forecasting models using deep learning. We compare the performances of fully-connected deep neural network (FC-DNN), deep long short-term memory (D-LSTM), and deep bidirectional long short-term memory (D-BLSTM) recurrent neural networks (RNNs). This allows us to examine the benefit of modeling dynamic patterns in emotion forecasting tasks. Our experimental results on the IEMOCAP benchmark dataset demonstrate that D-BLSTM and D-LSTM outperform FC-DNN by up to 2.96% in unweighted recall. When using both the current and past utterances, deep dynamic models show an improvement of up to 2.68% compared to their performance when using only the current utterance. We further analyze the benefit of using current and past utterance information compared to using the current and randomly chosen utterance information, and we find the performance improvement rises to 6.23%. The novelty in this study comes from its formulation of emotion forecasting problems and the understanding of how current and past audio-visual cues reveal future emotional information.**

I. INTRODUCTION

We define *emotion forecasting* as the task of predicting the future emotional state of a speaker from present and previous audio-visual behavioral cues. The research of emotion forecasting holds promise for its potential applications in a variety of domains, including human-computer or human-robot interactions [15]. For instance, an intelligent assistant system, such as Apple Siri [2] or Amazon Alexa [1], can benefit from emotion forecasting by forecasting a user's frustration. The early forecasting of the user's emotional state can help these systems to provide timely machine responses in advance to improve the overall user satisfaction. In this work, we explore deep learning–based static and temporal approaches for emotion forecasting tasks.

Emotion forecasting is an under-explored research area in affective computing. Traditional systems in affective computing mostly focus on emotion recognition, which is the task of identifying the current emotional state of a speaker. Traditional emotion recognition studies have shown that considering the temporal flow of emotion dynamics helps modeling the recognition system with improved accuracy [11], [16]. This leads to our first research question: **(Q1)** Can we capture the underlying salient information about the future emotional state better in the sequential pattern of current audio-visual cues than in static patterns? To address this question, we present a hypothesis: **(H1)** Emotion forecasting that considers sequential learning of information, or dynamic modeling, can enhance the forecasting performance compared to static modeling. In addition, previous research has indicated that the dynamics of emotion can occur in a longer range of time [20]. Hence, we present our second research question: **(Q2)** Can emotional history be helpful for forecasting emotion? To address this question, we introduce the second hypothesis: **(H2)** Analyzing temporal flow of audio-visual information from both the previous and present utterance (i.e., history-added forecasting) can ameliorate the emotion forecasting performance in comparison with using the current utterance only.

Prediction or forecasting of future events from current data has been studied in various fields, e.g., human activity forecasting [18], financial event prediction [19], and facial action event prediction [8]; however, it is less explored whether and how these methods can be used in emotion forecasting. For instance, Hoai and Torre addressed the problem of early detection of temporal facial action events from partial information [8]. Emotions and facial action events (e.g., smiling) are inherently different in that emotions do not have clear-cut boundaries as facial action events do (e.g., onset and offset of smiling). In emotion recognition, Kim and Provost [12] showed that only partial sub-utterance information with an estimated region of emotion saliency can achieve a comparable accuracy to the full utterance data. These studies have demonstrated the potential of emotion prediction models at the intra-utterance level; however, our work differs from these studies in its investigation of emotion forecasting at the inter-utterance level and its understanding of inter-utterance dynamical patterns.

In this paper, we use the dyadic conversations in the IEMOCAP database, which is segmented into variable-length utterances [4]. Table I provides a detailed overview of experiments to test **H1** and **H2**. The four experimental approaches are described below:

1) UF-cur is utterance forecasting using only the current utterance. The forecasting window is an utterance step, which means the number of speaking turns after the

current speaking turn (Section IV-B.1.a).
2) TF-cur is time forecasting using only the current utterance. The forecasting window is a time range (Section IV-B.1.b).
3) UF-his is utterance forecasting using both of the current and previous utterances (Section IV-B.2.b).
4) TF-his is time forecasting using both of the current and previous utterances.

We use the UF-cur and TF-cur approaches to test **H1**. As shown in Table II, we apply a fully connected deep neural network (FC-DNN) as a baseline static model. There are several dynamic modeling methods, including recurrent neural network (RNN) and its variants, and we choose to use deep long short-term memory (D-LSTM) and a deep bidirectional long short-term memory (D-BLSTM) network. **H2** is tested using UF-his and TF-his approaches. Comparing with the dynamic models of UF-cur and TF-cur, we explore the potential of adding previous history in forecasting emotion. We further perform an experiment with a randomly added utterance instead of the previous utterance, and we use this as a second baseline to demonstrate the effectiveness of a history-added forecasting technique.

Fig. 1 demonstrates an overview of the total process of emotion forecasting. First, we process the audio-visual data with different forecasting windows. Then, they are fed into several DNNs to test **H1** and **H2**. Our comparison of history-less and history-added emotion forecasting performance suggests the latter technique, which realizes the emotional flow from a prolonged history, is more successful at forecasting future emotion. Therefore, our work provides an insight that forecasting future emotion may depend on analyzing the temporal flow of not only the current but also the continued previous data.

TABLE I

EXPERIMENTAL APPROACHES FOR EVALUATION OF **H1** AND **H2**. THE UF-CUR AND TF-CUR APPROACHES ARE USED TO TEST **H1**, AND THE REMAINING APPROACHES ARE USED TO TEST **H2**.

Forecasting Window \ **History**	History-less (cur)	History-added (his)
Utterance Forecasting (UF)	UF-cur	UF-his
Time Forecasting (TF)	TF-cur	TF-his

TABLE II

PROPOSED AND BASELINE EXPERIMENTAL MODELS FOR TESTING **H1** AND **H2**.

Static Model	**Temporal Model**
FC-DNN	D-LSTM
	D-BLSTM

II. RELATED WORKS

In a related field of human action recognition, Davis and Tyagi [6] presented a probabilistic inference framework for a reliable human activity classification with the concept of using the least amount of temporal video information. A similar work was conducted by Ryoo [18], where he considered the sequential nature of human activity for early detection of unfinished action from the video sequences. He used a "dynamic bag of words" model, a probabilistic model which divides the activity model and the observed sequence into multiple segments to find the structural similarity between them. With a more generalist idea of early event detection, Hoai and Torre [8] used structured-Output Support Vector Machine (SOSVM) to introduce Max-Margin Early Event Detectors (MMED) to detect and localize an event before it was completed. Their method was proved to be faster and more reliable compared to the other complete event detectors such as Support Vector Machine (SVM) and Hidden Markov Model (HMM). The forecasting of future event was also investigated in financial data research and electrical load forecasting [19], [10]. In the field of emotion recognition, Kim and Provost [12] explored time and duration pattern of emotional utterances. They analyzed how only a portion of emotional data can acquire comparable accuracy to complete data, which drives us to hypothesize that emotional effect can stay longer than the specific utterance length. Wöllmer et al. [20] investigated the effect of including the past and future utterances for emotion recognition. In contrary to their work, we focus on forecasting, where we use the present and past temporal data to predict the future emotional state.

A closely related work was done by Noroozi et al. [17], where they predicted the sequence of emotional state from speech. In their work, they manually concatenated four emotional speech annotated as boredom, fear, joy and sadness, and thus they formulated a time series. Using a nonlinear autoregressive model, they predicted the next emotional group from the learned sequence. Our work differs from their work in that we have not made any fixed sequence or manual concatenation of emotional signals to predict the future emotion. Rather, we retain the conversation and forecast the next step from the current and past samples of data. Although forecasting actions or events from previous data were analyzed in many fields, predicting future emotion has not been explored much, with the exception of [17].

III. DATASET

IEMOCAP [4] contains approximately 12 hours of audio-visual data from five different pairs of actors of both genders. Ten actors were recorded in dyadic sessions in order to facilitate a more natural interaction and expression of the targeted emotion. The database contains audio data and facial motion capture (MoCap) marker trajectories. We use the speech data and 46 three-dimensional MoCap trajectories—captured over six face regions: chin, forehead, cheek, upper eyebrow, eyebrow, and mouth—similar to [16].

IEMOCAP is divided into 151 dialogs. The dialogs have an average length of five minutes, where only one of the speakers wears motion capture markers. The dialogs are further segmented into variable-length utterances, where each utterance has an average length of 4.77 ± 3.34 seconds. Every

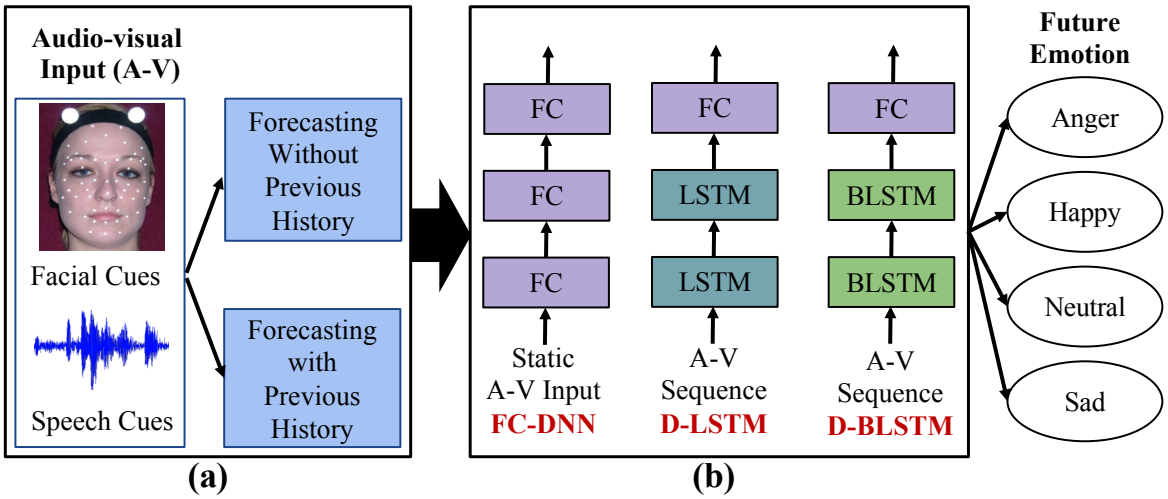

Fig. 1. An overview explaining the total process of emotion forecasting. First, in block **(a)**, the forecasting problem is formulated for the history-less and history-added technique using facial and speech features. In block **(b)**, various DNN models are used for forecasting. It is shown later with the history-less model, that dynamic models (D-LSTM/D-BLSTM) work better than static models (FC-DNN). Furthermore, the significance of adding the previous history while forecasting future emotion is also demonstrated. The facial figure is generated from [4].

utterance was annotated by at least three evaluators for the categorical and dimensional evaluation. There are $10,039$ utterances with ten different categories of emotion. To be consistent with previous work on IEMOCAP [16], [12], we use only four categorical emotions, namely, angry happy, neutral and sad. The average number of utterances over 10 speakers for each emotion are: 58.00 ± 25.41 for angry, 115.80 ± 27.85 for happy, 56.10 ± 22.03 for neutral and 62.50 ± 23.28 for sad.

IV. METHODOLOGY

A. Feature Extraction and Data Processing

Following the work of Metallinou et al. [16], both the audio and visual data were extracted at the same frame rate of 25 ms and the window of 50 ms.

For the audio data, we extract pitch, energy, 12 Mel Frequency Cepstral Coefficients (MFCC) and 27 Mel Filter Bank (MFB) coefficients, and we use Praat for feature extraction [3]. The 46 total 3-D markers allow for 138-dimensional facial landmark representation. Because of the noise in facial markers, some of the markers' output in some frames do not contain any number (NaN). We use linear interpolation method to remove the NaN values. If the total amount of NaN values in an utterance is greater than 30% of total frames, we exclude the utterance. Additionally, we exclude utterances with the audio data that has zero pitch in all frames since these utterances do not contain meaningful audio information.

For our baseline static model, we use utterance-level statistical features, namely, mean, standard deviation, first quantile, third quantile, and inter-quartile range. Therefore, building in the 179 frame-level features (41 audio and 138 facial landmark features), we have 895 statistical features in total. The features are normalized speaker-wise using z-normalization. D-LSTM and D-BLSTM use window-level features. The windows are made by taking statistical features over 30 frames with a 50% overlap. The features are z-normalized per windows for each speaker. As the IEMOCAP utterances have different length, we use zero-pad and later used a masking layer [5]. All the experiments were done using Keras [5].

B. Emotion Forecasting Methods

As stated before, we analyze emotion forecasting using four approaches. The approaches are divided based on the forecasting windows and presence of previous history utterance in forecasting.

1) Forecasting Windows: Due to the difference of length in various utterances of IEMOCAP database, we present two approaches of forecasting windows. The utterance forecasting will consider only the speaking turns after which forecasting will be done, while the time forecasting will consider the time amount for forecasting. Both the approaches are described below.

a) Utterance Forecasting (UF): For forecasting with utterance steps, we consider the emotional flow of one participant in a dialog only. For 1 utterance forecasting (UF

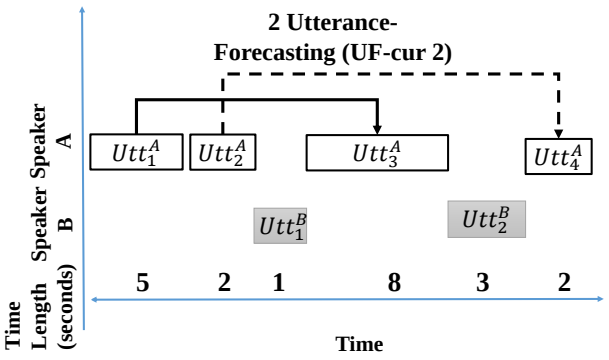

Fig. 2. An example of TF-cur 2. Here, the data of Utt_1^A will be used to forecast the emotion of the utterance Utt_3^A .

1), we choose the data of the present utterance and the label for the next utterance of the same speaker in the same dialog. Similarly, for k-utterance forecasting (UF k), the data will come from current utterance and the label will come from the utterance of n future steps.

To illustrate, let L be the set of all utterance labels of a particular speaker A, and U be the set of all utterance feature data of A. If A has n number of utterances, then $L = \{L_1, L_2, L_3 ... L_n\}$, and

$$U = \{Utt_1^A, Utt_2^A, Utt_3^A ... Utt_n^A\}$$

where L_x is the label of utterance Utt_x^A, where $x = 1, 2, 3 ... n$.

As in IEMOCAP, a dialog contains a complete scenario of emotional conversation, the forecasting has to be done within a dialog. For instance, the mapping of feature data to the labels for a k utterance forecasting will be $Utt_1^A \rightarrow L_{1+k}$ and so on. If the dialog has d number of utterances, then for the dth utterance, Utt_d^A, we cannot have the L_{d+k} labels. If $k > 1$, we will have more utterances in every dialog which have no labels and hence, must be discarded. Hence, the number of the utterances decreases in utterance forecasting (UF) 1, 2, and 3, as stated later in this section.

The process is described in Fig. 2, which depicts the dataset re-processing for UF 2. If in a dialog, the conversation alternate between speaker A and speaker B: $A \rightarrow B \rightarrow A \rightarrow B \rightarrow A$, then the utterance Utt_3^A's label would be used for the data of utterance Utt_1^A. For UF 0, which is just emotion recognition, we have 2924 utterances ; for UF 1, we have 2823 utterances; 2734 for UF 2 and 2662 for UF 3. These utterances have a mean time-distance (explained in the next section) from the current utterance to the target forecasting utterance of 4.71 ± 5.37 seconds 12.78 ± 8.55 seconds and 21.48 ± 10.78 seconds, respectively. Note that, for utterance Utt_1^A, the forecasting label has to pass one utterance of the other speaker (B), while utterance Utt_2^A has to pass two utterances of the other speaker. This creates inconsistency in the dataset and results in a large standard deviation of time-distance among the steps.

b) Time Forecasting (TF): Due to the inconsistency explained in previous section, we introduce the *time-distance* through which future emotion will be forecasted. Time-distance is defined as the time length from the current utterance to the forecasted utterance's emotional label for UF 1, 2 or 3 . Additionally, to calculate the time-distance, we also consider the length of the other speaker's utterance that falls in between the utterances of the same speaker. The calculation of time-distance starts from the middle point of the current utterance and ends at the middle point of the target utterance whose emotion is to be forecasted.

We divide the time-distances into several groups of 5-seconds range. TF 1 will have a forecasting time-distance of 1 second to 5 seconds, TF 2 will have a time-distance length of 5 seconds to 10 seconds, and TF 3 will have a time-distance of 10 seconds to 15 seconds. For example, in Fig. 1, the first forecasting will have a time-distance of 9.5 seconds $(2.5 + 2 + 1 + 4)$, and the second forecasting will have a time-distance of 14 seconds $(1 + 1 + 8 + 3 + 1)$. Therefore, although in terms of the forecasting window of UF, they will fall into same UF 2, in terms of TF, the first one will be part of TF 2 and the second one will be a part of TF 3. For TF 1, 2 and 3, we have 2006, 1763, and 1612 utterances. The mean time-distance for TF 1, 2 and 3 are 3.10 ± 1.01 seconds, 7.41 ± 1.49 seconds, and 12.60 ± 1.45 seconds.

2) Absence or Presence of Previous History:

a) History-Less Emotion Forecasting: In the history-less technique, we experiment emotion forecasting using the current audio-visual data only with both forecasting windows, UF and TF. Fig. 2 depicts the process of history-less emotion forecasting. As only current utterance is used for the task, we denote the history-less experimental approaches as UF-cur and TF-cur.

b) History-Added Emotion Forecasting: In this technique, in addition to the present emotional data, the prior utterance's data will also be used to forecast future emotion. An example is illustrated in Fig.3, where for 2 utterance forecasting (UF 2), along with the current utterance Utt_2^A, the previous utterance Utt_1^A is also taken into account and the features are concatenated. For instance, considering the illustration of Section IV-B.1.a, for a k utterance forecasting, the data will be,

$$U = \{Utt_1^A, Utt_{1,2}^A, Utt_{2,3}^A ... Utt_{n-k-1,n-k}^A\}$$

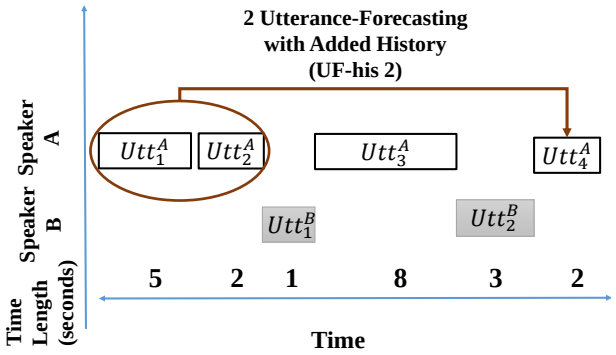

Fig. 3. An example of emotion forecasting with added history from the previous utterance. Here to forecast the emotion of Utt_4^A , the data of both Utt_1^A and Utt_2^A are used.

and the labels will be $L = \{L_{1+k}, L_{2+k}...L_n\}$

where

$$Utt^A_{n-k-1,n-k} = concatenation(Utt^A_{n-k-1}, Utt^A_{n-k})$$

We conduct the history-added emotion forecasting using UF-his and TF-his approaches.

C. Emotion Forecasting Model

1) Fully Connected Deep Neural Network (FC-DNN): Inspired by biological networks, FC-DNN is one of the most widely used tools for mapping static input to output. With a forward and backward propagation and a nonlinear activation function, FC-DNNs are widely used for capturing complex structure of information. In our work, we will use FC-DNN as a baseline model.

2) Deep Recurrent Neural Network (D-RNN): FC-DNN can capture the high-level structure in the data, however, it cannot identify or model dynamic properties in sequential data. RNN model has the ability to capture the temporal properties of the data. The LSTM network is an enhanced version of the RNN that has the added advantage of handling the vanishing gradient problem [9]. An LSTM layer is composed of recurrently connected memory blocks, where the memory cells have three gate units: the input, output, and forget gates. Briefly, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate [16]. Thus, information is retained and stored over an extensive time period. The internal cell state of an LSTM network (c^τ) at time τ is computed based on the current input (x^τ) and the previous cell state ($c^{\tau-1}$). The combination of input (i^τ) and forget (F^τ) gates control how much the previous cell state $c^{\tau-1}$ and the current input x^τ contribute to the current cell state c^τ. The activation function for both forget and input gates is a sigmoid function (σ) that outputs values between 0 and 1. The current cell state c_τ is determined by the following equations:

$$i^\tau = \sigma(W_{xi}x^\tau + W_{hi}h^{\tau-1} + W_{ci}c^{\tau-1} + b^i) \quad (1)$$
$$F^\tau = \sigma(W_{xf}x^\tau + W_{hf}h^{\tau-1} + W_{cf}c^{\tau-1} + b^f) \quad (2)$$
$$\tilde{c}^\tau = tanh(W_{xc}x^\tau + W_{hc}h^{\tau-1} + b^c) \quad (3)$$
$$c^\tau = F^\tau * c^{\tau-1} + i^\tau * \tilde{c}^\tau \quad (4)$$
$$o^\tau = \sigma(W_{xo}x^\tau + W_{ho}h^{\tau-1} + W_{co}c^{\tau-1} + b^o) \quad (5)$$
$$h^\tau = o^\tau * tanh(c^\tau) \quad (6)$$

where o^τ is the output gate that determines the contribution of current cell state c^τ to the current output h^τ. We can also rewrite h^τ and c^τ as follows:

$$(h^\tau, c^\tau) = \mathcal{G}(x^\tau, h^{\tau-1}, c^{\tau-1})$$

where $\mathcal{G}$ is the LSTM activation function. To process the input sequence in both forward and backward directions, the bidirectional LSTM (BLSTM) is used to capture the dynamics in both past and future samples. For BLSTM, h^τ will be composed of both the forward and backward directions, $h^\tau = [\overrightarrow{h^\tau}; \overleftarrow{h^\tau}]$ and it is defined as follows:

$$(\overrightarrow{h}^\tau, \overrightarrow{c}^\tau) = \overrightarrow{\mathcal{G}}(x^\tau, \overrightarrow{h}^{\tau-1}, \overrightarrow{c}^{\tau-1}) \quad (7)$$
$$(\overleftarrow{h}^\tau, \overleftarrow{c}^\tau) = \overleftarrow{\mathcal{G}}(x^\tau, \overleftarrow{h}^{\tau-1}, \overleftarrow{c}^{\tau-1}) \quad (8)$$

Our main architecture will contain deep LSTM (D-LSTM) / BLSTM (D-BLSTM) networks, followed by a FC-DNN.

D. Experimental Setup

Our FC-DNN model contains three fully connected layers with 256 memory units and one output softmax layer. The D-LSTM uses two LSTM layers with 128 memory units each. After that, an FC layer is used with RELU as activation and finally a softmax output layer is stacked with the network. The D-BLSTM is built up with similar architecture with the LSTM layers having 128 memory units in both direction. For all FC-DNN, D-LSTM and D-BLSTM, a rectified linear unit (RELU) is used as the activation unit. We use a learning rate of 0.0001 with the adam [13] optimizer and a batch size of 128. We also use an early stopping criteria, where the training stops if the validation accuracy does not improve after ten consecutive epochs.

To measure the performance, we calculate the Unweighted (Average) Recall (UAR), which is defined as the mean recall of all the emotion classes over the ten test speakers. As compared to the Weighted Recall (WR), UAR can represent unbiased performance. We perform leave-one-speaker-out cross-validation. For each of the ten test speakers, we choose

one speaker for validation at random and use the remaining eight speakers for training. The validation accuracy is used for choosing the number of epochs for early stopping. As suggested in previous research [7] , for testing the significance level of difference between different approaches, we use a paired t-test over ten speakers. We claim significance when $p<0.05$.

V. Results and Discussion

As stated in Section I, we use the UF-cur and TF-cur approaches for addressing **H1** and UF-his and TF-his for addressing **H2**.

A. History-Less Emotion Forecasting

Table III summarizes the results of UF-his and TF-his experiments. For UF-cur 1, 2, and 3, the results support **H1**. Although not statistically significant ($p>0.05$) , we achieve 0.97%, 1.28%, and 1.59% improvement using D-LSTM over FC-DNN, while we observe 0.97%, 1.91%, and 0.85% improvement using D-BLSTM. Nevertheless, we achieve a better set of accuracy with D-BLSTM than FC-DNN, it does not seem to work better than D-LSTM, particularly UF-cur 3. This may indicate that, as D-BLSTM has more complex modeling process than D-LSTM and it needs a large amount of data as well, our data quantity may be insufficient (2662 utterances for UF 3) for that. Moreover, while we consider UF, there can be a varying number of utterances in between two utterances of the same speaker as explained in Section IV-B.1.a. Thus, the time-distances have large standard deviations (10.78 seconds for UF 3), which can affect the forecasting performance.

TABLE III
UAR(%) comparison between static (FC-DNN) and dynamic (D-LSTM and D-BLSTM) models for both UF-cur and TF-cur approaches. The sign "[*]" denotes that the model is statistically significant ($p<0.05$) compared to the FC-DNN

Forecasting Window	FC-DNN	D-LSTM	D-BLSTM
UF-cur 1	57.96	58.93	58.95
UF-cur 2	54.45	55.73	56.36
UF-cur 3	51.40	52.99	52.25
TF-cur 1	57.29	59.72	59.88 [*]
TF-cur 2	53.59	55.99	56.55 [*]
TF-cur 3	52.65	54.41 [*]	55.22

Table III shows the performance comparison for TF-cur 1, 2, and 3. It shows that D-LSTM performs better than the FC-DNN model by 1.43%, 2.40%, and 1.76% ($p<0.05$) for TF-cur 1, 2 and 3 respectively. When we consider both forward and backward dynamics by taking into account a bidirectional temporal model (D-BLSTM), we find better accuracy than FC-DNN model by 1.59% ($p<0.05$), 2.96% ($p<0.05$), and 2.57% respectively. Therefore, the above experiments supports our hypothesis **H1** that temporal model can perform better than the static model.

B. History-Added Emotion Forecasting

As stated previously, adding history to present information incorporates more temporal contexts, which can result in an enhanced forecasting performance. Table IV describes the results of both UF and TF approaches with the history of previous utterance added. Adding history improved the D-LSTM accuracy for UF-his 1, 2, and 3 by 1.75%, 0.78%, and 0.34% compared to UF-cur D-LSTM approaches. We observe even more improvement in D-BLSTM performance. D-BLSTM is improved by 2.38% ($p<0.05$), 2.01%, and 2.39% ($p<0.05$) for UF-his 1, 2, and 3.

For the TF-his approach, both the D-LSTM and D-BLSTM performance improves over the history-less performance. We observe an improvement of 0.64%, 0.21%, and 1.63% ($p<0.05$) for D-LSTM and 0.93%, 0.10%, and 2.68% for D-BLSTM in TF-his 1, 2, and 3 than the TF-cur approaches. Moreover, Table IV shows that in every case, unlike UF-cur result, the D-BLSTM outperforms D-LSTM performance. This may imply that D-BLSTM can process the information in a better way when adequate history is provided.

TABLE IV
UAR(%) results of emotion forecasting with added history information: the performance of dynamic modeling with UF-his and TF-his experimental approaches in comparison with the dynamic modeling of UF-cur and TF-cur approaches. The sign "[*]" denotes that the model with history-added technique is statistically significant ($p<0.05$) compared to the history-less technique of the model.

Forecasting Window	D-LSTM	D-BLSTM
UF-his 1	60.68	61.33 [*]
UF-cur 1	58.93	58.95
UF-his 2	56.51	58.37
UF-cur 2	55.73	56.36
UF-his 3	53.33	54.64 [*]
UF-cur 3	52.99	52.25
TF-his 1	60.36	60.81
TF-cur 1	59.72	59.88
TF-his 2	56.20	56.65
TF-cur 2	55.99	56.55
TF-his 3	56.04 [*]	57.90
TF-cur 3	54.41	55.22

Furthermore, instead of previous utterance, we add a

randomly chosen utterance of the same speaker (i.e., UF-random and TF-random) to compare the effect of adding random data with current utterance rather than the history. Table V shows that for every case, compared to UF-random or TF-random, UF-his and TF-his performs significantly better ($p<0.05$). The UF-his 1, 2, and 3 has an improvement of 5.58%, 4.50%, and 4.65% respectively with D-LSTM and 4.69%, 5.64%, and 5.30% respectively with D-BLSTM. Similarly, for TF-his 1, 2, and 3, the improvement is 5.32%, 4.86%, and 4.94% respectively with D-LSTM and 3.79%, 4.35%, and 6.23% respectively with D-BLSTM, comparing with the randomly added data. As adding random utterance in the data may harm the emotional temporal flow, we observe a diminished UAR performance. It implies that our improvement in history-added D-BLSTM performance is not merely coming from adding extra information, rather, by taking the proper history context into account. Therefore, we show that adding history information from one previous utterance can add an emotional context for the network to see the temporal flow pattern of the utterances and predict the future emotion, and hence, the results support **H2**.

The reason for D-BLSTM network's superior performance in both UF-his and TF-his experiments may be that, as a more complex modeling framework compared to the D-LSTM, the D-BLSTM model can see both past and future information. Thus, by achieving the history information, it can achieve substantial information by moving in both directions.

TABLE V
UAR (%) COMPARISON OF UF-HIS/TF-HIS WITH UF-RANDOM/TF-RANDOM. THE SIGN "[*]" DENOTES THE STATISTICALLY SIGNIFICANT ($p<0.05$) RESULT OF HISTORY ADDED TECHNIQUE, COMPARED TO THE RANDOMLY ADDED DATA.

Forecasting Window	D-LSTM	D-BLSTM
UF-his 1	60.68 [*]	61.33 [*]
UF-random 1	55.10	56.62
UF-his 2	56.51 [*]	58.37 [*]
UF-random 2	52.01	52.73
UF-his 3	53.33 [*]	54.64 [*]
UF-random 3	48.68	49.34
TF-his 1	60.36 [*]	60.81 [*]
TF-random 1	55.04	57.02
TF-his 2	56.20 [*]	56.65 [*]
TF-random 2	51.34	52.30
TF-his 3	56.04 [*]	57.90 [*]
TF-random 3	51.10	51.67

C. Further Analysis

As shown in Table VI and VII, we further investigate the performance of utterances that have the same target forecasting and current emotion labels. We examine whether the performance of emotion forecasting is biased toward that of emotion recognition. We define the term *same-label* as, for any forecasting approach, the ratio of instances, where forecasting and recognition ground truths are same, to all the forecasting ground truth labels. Table VI demonstrates that in all forecasting approaches, more than two-third of the forecasting utterances have the same ground truth labels with the recognition. Table VII shows an example of the forecasting performance, when we have such identical labels. We observe the highest UAR with the 'Happy' emotion (83.25%), and the lowest with the 'Neutral' emotion (36.55%). However, When the forecasting label is different from current emotion label, we achieve lower UAR as expected. For instance, UF-cur 3 D-BLSTM results achieve up to 36.72% for angry, 41.30% for happy, 17.09% for neutral, and 32.10% for sad classes. Hence, the results imply that the training and testing data used for the forecasting experiments may be biased towards the current emotion labels. Therefore, an extensive analysis on forecasting window is needed to develop a more robust emotion forecasting system.

TABLE VI
THE RATIO OF IDENTICAL LABELS OF FORECASTING AND RECOGNITION, TO ALL THE FORECASTING LABELS FOR DIFFERENT FORECASTING WINDOWS (I.E., SAME-LABELS).

Forecasting Window	Same-labels
UF 1	0.73
UF 2	0.68
UF 3	0.64
TF 1	0.74
TF 2	0.71
TF 3	0.67

TABLE VII
AN EXAMPLE FOR THE PER-CLASS RECALL SCORE (%), WHEN THE FORECASTING LABEL IS SAME WITH CURRENT EMOTIONAL LABEL. THE EXAMPLE IS TAKEN FROM UF-CUR 3 D-BLSTM RESULT. 'GT' REFERS TO FORECASTING GROUND TRUTH AND 'PREDICT' REFERS TO FORECASTING LABELS PREDICTED BY THE NETWORK.

Predict \ GT	Angry	Happy	Neutral	Sad
Angry	**65.66**	19.25	12.08	3.02
Happy	6.72	**83.25**	6.60	3.42
Neutral	28.99	25.63	**36.55**	8.82
Sad	7.44	23.67	5.85	**63.03**

VI. CONCLUSIONS

This paper describes the formulation of emotion forecasting techniques from current and past audio-visual data. For UF-cur and TF-cur approaches, we demonstrate the dynamic models (D-LSTM and D-BLSTM) always outperform the static model (FC-DNN). For the history added approaches (i.e., UF-his and TF-his), we use the D-LSTM and D-BLSTM models and our findings show an enhanced forecasting performance in all the experiments comparing with the history-less technique (i.e., UF-cur and TF-cur). We further experiment with an added random utterance instead of the history, and the performance declined significantly, indicating the effectiveness of adding history information in emotion forecasting. The experimental results for testing **H1** and **H2** lead us to conclude that emotion forecasting can improve considerably if we forecast with a deep bidirectional dynamic model (D-BLSTM) and take a previous utterance along with the present utterance.

One of the limitations of our work is that we have not tested if our proposed models are optimized for the emotion forecasting task. Additionally, the deep learning framework can be sensitive and unstable. An extensive hyperparameter search can make the network less sensitive. Hence, investigating an optimized model with a comparatively stable framework would be our next research step. There can be several future research directions from this work. First, we focus only on the emotional flow of one speaker and do not consider the influence of the other speaker, while previous studies showed improvement of accuracy in emotion recognition, when the mutual influence was taken into account [14]. Hence, forecasting performance can be enhanced considerably, if the effect of other person's emotion is incorporated in the modeling process. Second, emotion can be highly contextual, which means emotional flow can be biased in different scenarios. Therefore, forecasting result can be improved by considering the context.

REFERENCES

[1] *Amazon Alexa*, accessed September 22, 2018. `https://developer.amazon.com/alexa`.

[2] *Siri*, accessed September 22, 2018. `https://www.apple.com/siri/`.

[3] Paul Boersma and David Weenink. Praat: Doing phonetics by computer (version 5.3.51). 01 2007.

[4] C. Lee A. Kazemzadeh E. Mower S. Kim J. Chang S. Lee C. Busso, M. Bulut and S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, Dec 2015.

[5] François Chollet et al. Keras. `https://keras.io`, 2015.

[6] James W. Davis and Ambrish Tyagi. Minimal-latency human action recognition using reliable-inference. *Image Vision Comput.*, 24(5):455–472, May 2006.

[7] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, Oct 1998.

[8] M. Hoai and F. De la Torre. Max-margin early event detectors. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2863–2870, June 2012.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Nov 1997.

[10] S. J. Kiartzis, C. E. Zoumas, J. B. Theocharis, A. G. Bakirtzis, and V. Petridis. Short-term load forecasting in an autonomous power system using artificial neural networks. *IEEE Transactions on Power Systems*, 12(4):1591–1596, Nov 1997.

[11] Y. Kim and E. M. Provost. Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3677–3681, May 2013.

[12] Yelin Kim and Emily Mower Provost. Emotion spotting: discovering regions of evidence in audio-visual emotion expressions. In *ICMI*, 2016.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[14] Chi-Chun Lee, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Modeling Mutual Influence of Interlocutor Emotion States in Dyadic Spoken Interactions. In *Proceedings of Interspeech 2009*, Brighton, UK, September 2009.

[15] Dongkeon Lee, Kyo-Joong Oh, and Ho-Jin Choi. The chatbot feels you - a counseling service using emotional response generation. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 437–440, Feb 2017.

[16] A. Metallinou, A. Katsamanis, M. Wöllmer, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification (extended abstract). In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 463–469, Sept 2015.

[17] F. Noroozi, N. Akrami, and G. Anbarjafari. Speech-based emotion recognition and next reaction prediction. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, May 2017.

[18] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 1036–1043, Washington, DC, USA, 2011. IEEE Computer Society.

[19] Francis E. H. Tay and Lijuan Cao. Application of support vector machines in financial time series forecasting. *Omega*, 29(4):309–317, 2001.

[20] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan. Analyzing the memory of blstm neural networks for enhanced emotion classification in dyadic spoken interactions. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4157–4160, March 2012.