Multi-modal Conditional Attention Fusion for Dimensional Emotion Prediction

Shizhe Chen, Qin Jin*
School of Information, Renmin University of China
{cszhe1, qjin}@ruc.edu.cn

ABSTRACT

Continuous dimensional emotion prediction is a challenging task where the fusion of various modalities usually achieves state-of-the-art performance such as early fusion or late fusion. In this paper, we propose a novel multi-modal fusion strategy named conditional attention fusion, which can dynamically pay attention to different modalities at each time step. Long-short term memory recurrent neural networks (LSTM-RNN) is applied as the basic uni-modality model to capture long time dependencies. The weights assigned to different modalities are automatically decided by the current input features and recent history information rather than being fixed at any kinds of situation. Our experimental results on a benchmark dataset AVEC2015 show the effectiveness of our method which outperforms several common fusion strategies for valence prediction.

Keywords

Continuous dimensional emotion prediction; Multi-modal Fusion; LSTM-RNN

1. INTRODUCTION

Understanding human emotions is a key component to improve human-computer interactions [1]. A wide range of applications can benefit from emotion recognition such as customer call center, computer tutoring systems and mental health diagnoses.

Dimensional emotion is one of the most popular computing models for emotion recognition [2]. It maps an emotion state into a point in a continuous space. Typically the space consists of three dimensions: arousal (a measure of affective activation), valence (a measure of pleasure) and dominance (a measure of power or control). This representation can express natural, subtle and complicated emotions. There have been many research works on dimensional emotion analysis for better understanding human emotions in recent years [3, 4, 5].

Since emotions are conveyed through various human behaviours, past works have utilized a broad range of modalities for emotion recognition including speech [6], text [7], facial expression [8], gesture [9], physiological signals [10], etc. Among them, facial expression and speech are the most common channels to transmit human emotions. It is beneficial to use multiple modalities for emotion recognition.

Fusion strategies for different modalities in previous works can be divided into 3 categories, namely feature-level (early) fusion, decision-level (late) fusion and model-level fusion [11]. Early fusion uses the concatenated features from different modalities as input features for classifiers. It has been widely used in the literature to successfully improve performance [12]. However, it suffers from the curse of dimensionality. Also it's not very useful when features are not synchronized in time. Late fusion eliminates some disadvantages of early fusion. It combines the predictions of different modalities and trains a second level model such as RVM [13], BLSTM [14]. But it ignores interactions and correlations between different modality features. Model-level fusion is a compromise between the two extremes. The implementation of model-level fusion depends on the specific classifiers. For example, for neural networks, model-level fusion could be concatenation of different hidden layers from different modalities [15]. For kernel classifiers, model-level fusion could be kernel fusion [16]. As for Hidden Markov Model (HMM) classifiers, novel forms of feature interactions have been proposed [17].

In this paper, we propose a novel architecture for the fusion of different modalities called conditional attention fusion. We use Long-short term memory recurrent neural networks (LSTMs) as the basic model for each uni-modality since LSTMs are able to capture long time dependencies. For each time step, the fusion model learns how much of attentions it should put on each modality conditioning on its current input multi-modal features and recent history information. This approach is similar to human perceptions since humans can dynamically focus on more obvious and trustful modalities to understand emotions. Unlike early fusion, we dynamically combine predictions of different modalities, which avoids the curse of dimensionality and synchronization between different features. And unlike late fusion, the input features are interacted in a higher level to learn the current attention instead of being isolated without any interactions among different modalities. The main architecture is shown in Figure 2. We use the AVEC2015 dimensional

*Qin Jin is the corresponding author

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967286

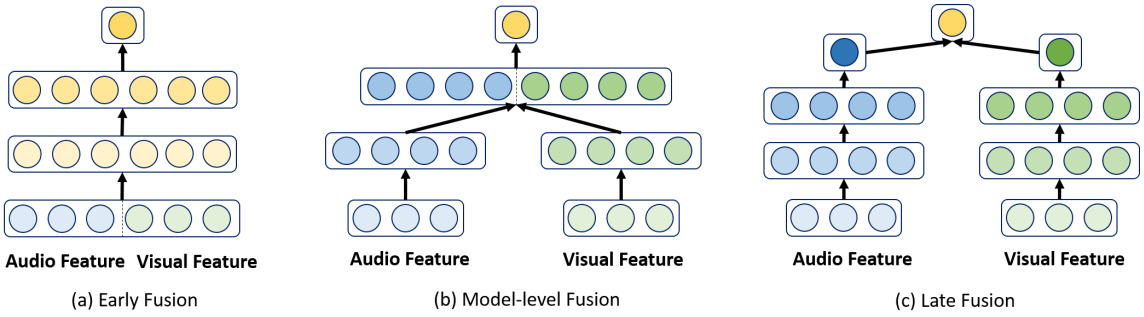

Figure 1: Three typical multi-modality fusion strategies. (a) Early fusion: concatenation of features from different modalities. (b) Model-level fusion: concatenation of high level feature representations from different modalities. (c) Late fusion: fusion of predictions from different modalities.

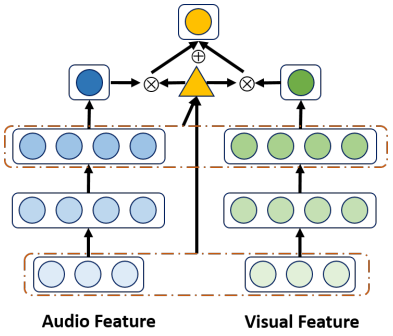

Figure 2: Conditional attention fusion model. The units in the dashed red line is concatenated together as the input of the triangle to learn the attention weights. $\otimes$ and $\oplus$ denote the element multiplication and addition respectively.

emotion dataset [5] to evaluate our methods. The results shows the effectiveness of our new fusion architecture.

2. MULTI-MODAL FEATURES

2.1 Audio Features

We utilize the OpenSMILE toolkit [18] to extract low-level features including MFCCs, loudness, F0, jitter and shimmer. All the features are extracted using 40ms frame window size without overlap to match with the groundtruth labels since it is demonstrated in [19] that short-time features can reveal more details and thus boost performance for affective prediction using LSTMs. The low-level acoustic features are in 76 dimensions.

2.2 Visual Features

Two sets of visual features are extracted from facial expression: appearance-based features and geometric-based features [5]. The appearance-based features are computed by using Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) and are compressed to 84 dimensions via PCA. The geometric-based features in 316 dimensions are computed from 49 facial landmarks. Frames where no face is detected are filled with zeros. We concatenate appearance-based features and geometric-based features as our visual feature representations.

3. EMOTION PREDICTION MODEL

3.1 Uni-Modality Prediction Model

Long short term memory (LSTM) architecture [20] is the state-of-the-art model for sequence analysis and can exploit long range dependencies in the data. In this paper, we use the peephole LSTM version proposed by Graves [21]. The function of hidden cells and gates are defined as follows.

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ c_t &= f_t \cdot c_{t-1} + i_t \cdot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \\ h_t &= o_t \cdot tanh(c_t) \end{aligned} \quad (1)$$

where i, f, o and c refers to the input gate, forget gate, output gate, and cell state respectively. $\sigma(\cdot)$ is the sigmoid function and $tanh(\cdot)$ is the tangent function.

3.2 Conditional Attention Fusion Model

Let x_t^a and x_t^v refer to the audio features and visual features respectively at the t^{th} frame. h_t^a and h_t^v refer to the outputs of the last hidden layer from uni-modality model with audio or visual features respectively. f_{θ_a} and f_{θ_v} refer to the uni-modality model which maps the audio or visual features into predictions. We define the conditional attention fusion of the predictions from the two modalities at timestep t as:

$$\widehat{y_t} = \lambda_t \cdot f_{\theta_a}(x_t^a, h_{t-1}^a) + (1 - \lambda_t) \cdot f_{\theta_v}(x_t^v, h_{t-1}^v) \quad (2)$$

$$\lambda_t = \sigma(W_g[h_t^a|h_t^v|x_t^a|x_t^v]) \quad (3)$$

where $[h_t^a|h_t^v|x_t^a|x_t^v]$ is the concatenation of the representations inside the bracket.

The λ_t is calculated based on the current audio and visual features and their high-level history information for two reasons. Firstly, the current input features are the most direct indicators to show whether the modality is reliable. For example, for facial features, inputs filled with 0s suggest that the current face detection fails and thus should be assigned with less confidence. Secondly, the weights assigned to each modality would be smoothed by considering high-level history features h_t^v and h_t^a in addition to current input features. In this way, the model can dynamically pay

attention to different modalities, which could improve the stability in different situations.

3.3 Model Learning

Intuitively, the acoustic features are more reliable when the acoustic energy is higher, because the headset microphone can record speech from both the subject speaker and other speakers in conversations. Higher energy may refers to higher confidence that the speech is from the target subject. Similarly the facial features are reliable only when faces are correctly detected. So adding such side information might be beneficial to learn the attention weights.

We transform the acoustic energy into scale [0, 1], and we use g_t^a to denote its value at the t^{th} timestep. For visual features, we use $g_t^v \in \{0, 1\}$ to indicate whether the subject's face is detected since the face detection provided in the dataset has no detection confidence. We therefore define the final loss function for one sequence as follows:

$$L_t^g = \frac{1}{2}(\alpha(g_t^a - \lambda_t)^2 + \beta(g_t^v - (1 - \lambda_t))^2) \quad (4)$$

$$L = \sum_t \frac{1}{2}(\widehat{y}_t - y_t)^2 + L_t^g \quad (5)$$

where α and β are hyper-parameters and are optimized on the development set. In practice, α and β are usually set to small values around 10^{-2} to avoid L_t^g over-affecting on λ_t.

The derivative of L_t^g with respect to λ_t is as follows:

$$\frac{\partial L_t^g}{\partial \lambda_t} = \beta g_t^v - \alpha g_t^a - \beta + (\alpha + \beta)\lambda_t \quad (6)$$

When g_t^a is high and g_t^v is low, (6) is close to $(\alpha+\beta)(\lambda_t - 1)$ (the extreme case when $g_t^a = 1$ and $g_t^v = 0$). The derivative is less than 0, which will push λ_t to increase to give acoustic features more confidence. But when $g_t^a \to 0$ and $g_t^v \to 1$, (6) is close to $(\alpha+\beta)\lambda_t$, which is larger than 0 and will push λ_t to decrease to focus on visual features. When $g_t^a \approx g_t^v$, the absolute value of the derivative would be small and thus L_t^g has less influence on λ_t.

4. EXPERIMENTS

4.1 Dataset

The AVEC2015 dimensional emotion dataset is a subset of the RECOLA dataset [22], a multimodal corpus of remote and collaborative affective interactions. There are 27 subjects in the dataset and are equally divided into training, development and testing sets. Audio, video and physiological data are collected for each participant for the first 5 minutes of interactions. Arousal and valence are annotated in scale [-1, 1] for every 40ms [5]. Since the submission times on testing set are limited, we carry out cross validation on the development set. We randomly select 5 subjects as the development set to optimize hyper parameters and the remained 4 speakers are used as the test set. We do the experiments 8 times. The concordance correlation coefficient (CCC) [5] works as the evaluation metric.

4.2 Experimental Setup

Annotation delay compensation [13] is applied because there exists a delay between signal content and groundtruth labels due to annotators' perceptual processing. We drop first N groundtruth labels and last N input feature frames.

Table 1: CCC performance of uni-modal features

	audio feature	visual feature
arousal	0.787	0.432
valence	0.595	0.620

Table 2: CCC performance of different loss functions on valence prediction

	mean	std
without L_t^g, $\alpha = \beta = 0$	0.672	0.046
with L_t^g, $\alpha = 0.04, \beta = 0.02$	0.684	0.041

N is optimized by non-temporal regression model SVR on training set. In this paper, N is optimized to be 20 frames for both audio and visual features. When predicting the result, the outputs of the model are shifted back by N frames. The missing predictions in the first N frames are filled with zeros. Finally, a binomial filter is applied to smooth the predictions. Annotation delay compensation and smoothing is applied in all the following experiments.

The input features are normalized into the range [-1,1]. For acoustic features, the LSTM has 2 layers and 100 cells for each layer. For visual features, the LSTM has 2 layers and 120 cells for each layer. The size of mini-batch is 256 and truncated backpropagation through time (BPTT) [23] is applied. The initial learning rate is set to be 0.01 with learning rate decay. Dropout is used as regularization. The training epochs are 100 and the model that achieves the best performance in development set is used as the final model.

We compare the conditional attention fusion model with early fusion, late fusion and model-level fusion. For early fusion, the LSTM has 150 units each layer, which has the similar size of parameters to other fusion methods. For late, model-level and conditional attention fusion, the parameters in LSTM are initialized with the trained uni-modal LSTMs. In order to avoid overfitting, we only fine-tune the network for 10 epochs with smaller initial learning rate 0.001. The hyper-parameters α and β are set to zeros for arousal prediction and 0.04, 0.02 respectively for valence prediction.

4.3 Experimental Results

Table 1 shows the prediction performance using uni-modality features. Acoustic features achieve the best performance on

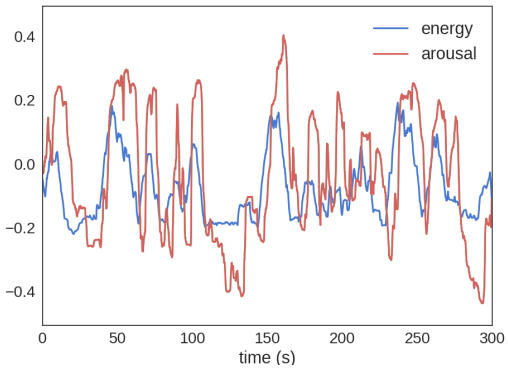

Figure 3: The relationship between arousal labels and acoustic energy on an example from dev set

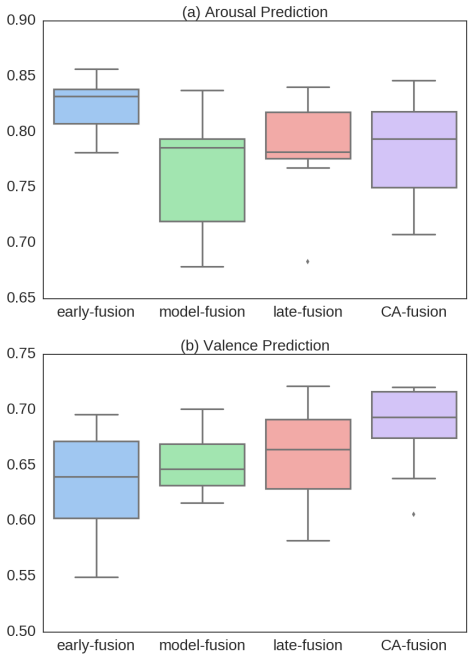

Figure 4: Performance of different fusion strategies

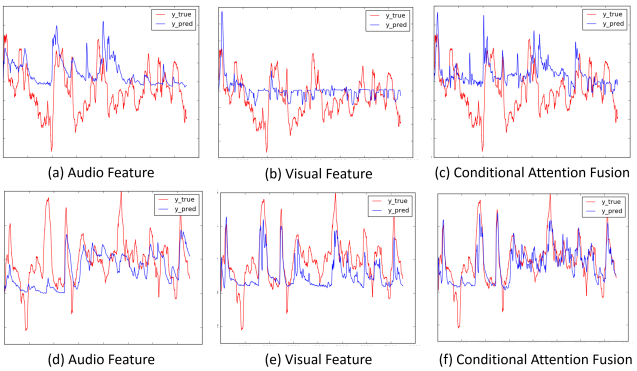

Figure 5: Examples of valence prediction on dev set. The red line is the groundtruth label and the blue line is the prediction.

Table 3: Valence prediction performance on test set

	rmse	pcc	ccc
Chen et al. [19]	0.111	0.59	0.567
Chao et al. [24]	0.103	0.627	0.618
CA-fusion	0.090	0.716	0.664

the arousal prediction and visual features are slightly better than acoustic features on valence prediction.

The performance of different fusion methods on arousal prediction is shown in Figure 4(a). Early fusion achieves the average best performance and our proposed fusion method performs the second best among all the fusion strategies. However, there is no significant difference between early fusion prediction and acoustic uni-modality prediction comparing Figure 4(a) with Table 1 (Student t-test with p-value = 0.07). We find that there exists a strong correlation between arousal and the acoustic energy, as shown in Figure 3 where we smooth the acoustic energy with window 100 and shift and scale it according to the mean and standard deviation between energy and arousal labels. And their Pearson Correlation Coefficient on development set is high to 0.558 and CCC is 0.4. This suggests that humans' perception of arousal may mainly base on acoustic features so fusing other modalities may bring less benefit.

But for valence prediction, all the fusion strategies outperform the original uni-modality models (as shown in Table 1). An interesting finding is that the higher level the fusion strategy applies, the better performance is achieved as shown in Figure 4(b). Among them, our proposed conditional attention fusion model achieves the best performance and significantly surpasses other fusion strategies by t-test ($p<0.02$ compared with the second best fusion strategy late fusion, and $p<0.007$ compared with others). This indicates that dynamically adapting fusion weights for different modalities is beneficial.

Table 2 shows the CCC performance with and without L_t^g in loss function. We can see that considering L_t^g in loss function can further improve performance since it helps to guide the importance of different modalities. It is might because of the insufficiency of data that the model is unable to learn the attention weights effectively without any supervised information. We also observe in our experiments that when α and β are around 10^{-2}, there is no significant change in prediction performance.

Figure 5 shows some examples of the emotion predictions from the conditional attention fusion method. The upper row shows the case where most of the visual features are missing and the bottom row is another case where the visual features can be extracted in most frames. We can see that the fusion method can make use of the complementary information automatically from audio and visual features in these two situations.

The valence prediction performance of the conditional attention fusion method on testing set is shown in Table 3. Chen et al. [19] use the same feature set as ours and Chao et al. [24] use more features including CNNs for valence prediction. The comparison further demonstrates the effectiveness of the conditional attention fusion method.

5. CONCLUSIONS

In this paper we propose a multi-modal fusion strategy named conditional attention fusion for continuous dimensional emotion prediction based on LSTM-RNN. It can dynamically pay attention to different modalities according to the current modality features and history information, which increases the model's stability. Experiments on benchmark dataset AVEC 2015 show that our proposed fusion approach significantly outperform the other common fusion approaches including early fusion, model-level fusion and late fusion for valence prediction. In the future, we will use more features from different modalities and apply strategies to express the correlation and independence of different modality features better.

6. ACKNOWLEDGEMENTS

This work is supported by National Key Research and Development Plan under Grant No. 2016YFB1001202.

7. REFERENCES

[1] Rosalind W. Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1-2):55–64, 2003.

[2] Stacy Marsella and Jonathan Gratch. Computationally modeling human emotion. *Communications of the ACM*, 57(12):56–67, 2014.

[3] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, 2013.

[4] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[5] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. Av+ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *International Workshop on Audio/visual Emotion Challenge*, pages 3–8, 2015.

[6] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[7] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[8] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[9] Stefano Piana, Alessandra Stagliano, Francesca Odone, Alessandro Verri, and Antonio Camurri. Real-time automatic emotion recognition from body gestures. *arXiv preprint arXiv:1402.5047*, 2014.

[10] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, 129:94–106, 2014.

[11] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing*, 3:e12, 2014.

[12] Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Aravind Namandi Vembu, and Rohit Prasad. Emotion recognition using acoustic and lexical features. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pages 366–369, 2012.

[13] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, Phu Le, Vidhyasaharan Sethu, and Julien Epps. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In *The International Workshop on Audio/visual Emotion Challenge*, 2015.

[14] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC 2015, Brisbane, Australia, October 26, 2015*, pages 73–80, 2015.

[15] Zuxuan Wu, Yu-Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pages 167–176, 2014.

[16] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turkey, November 12-16, 2014*, pages 508–513, 2014.

[17] Kun Lu and Yunde Jia. Audio-visual emotion recognition with boosted coupled HMM. In *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, November 11-15, 2012*, pages 1148–1151, 2012.

[18] Florian Eyben, Martin llmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. *Acm Mm*, pages 1459–1462, 2010.

[19] Shizhe Chen and Qin Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC 2015, Brisbane, Australia, October 26, 2015*, pages 49–56, 2015.

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

[21] Alex Graves. Generating sequences with recurrent neural networks. *Eprint Arxiv*, 2013.

[22] Fabien Ringeval, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013, Shanghai, China, 22-26 April, 2013*, pages 1–8, 2013.

[23] P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[24] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC 2015, Brisbane, Australia, October 26, 2015*, pages 65–72, 2015.