HIERARCHICAL TWO-LEVEL MODELLING OF EMOTIONAL STATES IN SPOKEN DIALOG SYSTEMS

Oxana Verkholyak[1], Dmitrii Fedotov[2], Heysem Kaya[3], Yang Zhang[4] and Alexey Karpov[1]

[1]St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia
[2] Institute of Communications Engineering, Ulm University, Ulm, Germany
[3]Department of Computer Engineering, Namık Kemal University, Çorlu, Tekirdağ, Turkey
[4] Huawei Noah's Ark Lab, Huawei Technologies, Shenzhen, China
overkholyak@gmail.com, dmitrii.fedotov@uni-ulm.de, hkaya@nku.edu.tr,
zhangyang86@huawei.com, karpov@iias.spb.su

ABSTRACT

Emotions occur in complex social interactions, and thus processing of isolated utterances may not be sufficient to grasp the nature of underlying emotional states. Dialog speech provides useful information about context that explains nuances of emotions and their transitions. Context can be defined on different levels; this paper proposes a hierarchical context modelling approach based on RNN-LSTM architecture, which models acoustical context on the frame level and partner's emotional context on the dialog level. The method is proved effective together with cross-corpus training setup and domain adaptation technique in a set of speaker independent cross-validation experiments on IEMOCAP corpus for three levels of activation and valence classification. As a result, the state-of-the-art on this corpus is advanced for both dimensions using only acoustic modality.

Index Terms— Emotion recognition, cross-corpus, context modelling, dialog systems, LSTM

1. INTRODUCTION

The field of automatic emotion recognition in spoken language advanced so far from simple classification of laboratory-controlled emotions to analysis of spontaneously expressed affects. Because emotions are usually considered in the context of social interactions, in addition to analyzing emotional speech of a single speaker, a significant research focus was directed at modelling context and mutual influence of both interlocutors in a dialog. Importance of analyzing context in the field of emotion recognition was previously shown by several authors: accuracy was improved with additional dialog-specific features, such as width and depth of the dialog, gender [1] and the dialog act of speaker turns [2].

Much work was dedicated to analyzing acoustical context on a frame level. In particular, Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) were

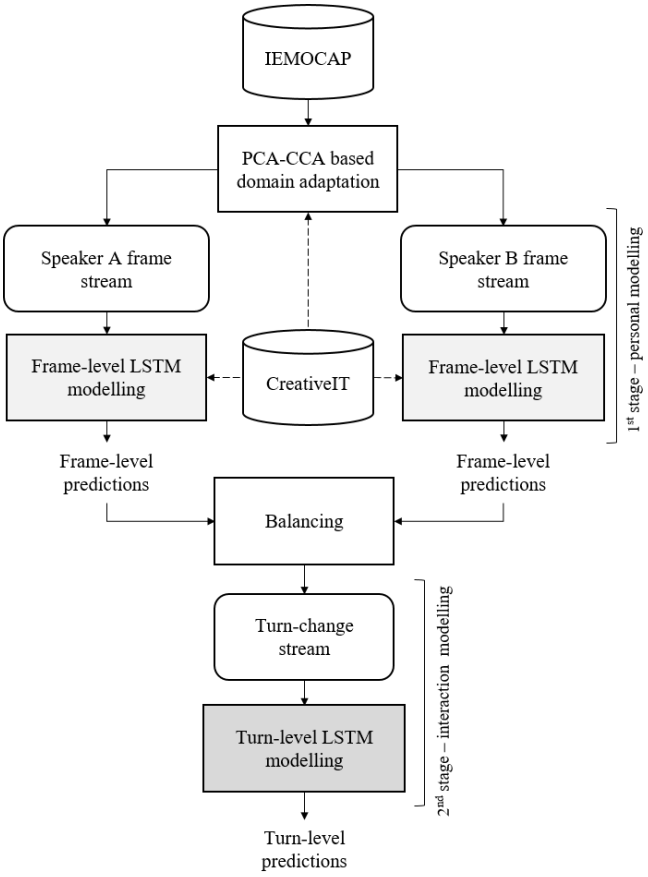

Fig. 1. Pipeline of the proposed framework

effectively used to track changes over the course of speech act [3, 4, 5]. In the light of these findings, we propose a two-level hierarchical system based on RNN-LSTM model that takes advantage of both acoustical context of a speaker on the frame level, as well as his interlocutor's emotional flow and gender information on the turn level to predict valence and activation primitives over the course of entire dialog. All experiments are conducted on the IEMOCAP [6] emotional dialog database.

2. PROPOSED METHOD

The proposed method is implemented in two stages as depicted in Figure 1. In the first stage, the personal behavior of each speaker is modelled independently via frame-level LSTM neural network trained on a database with continuous annotation. Due to lack of continuous labels in IEMOCAP corpus, another database, namely CreativeIT [7], is used for frame-level modelling. Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) based domain adaptation technique as proposed in [8] is performed to cope with the mismatched domain condition.

In the second stage, the interaction in a conversation between two speakers is modelled via turn-level LSTM trained on a dialog database with suitable annotation. The transition between two stages happens in a balancing operation, which transforms speakers' individual frame-level prediction streams into dialog-level turn-change stream. The experiments for activation and valence are conducted separately. The detailed description of each step is given in the following subsections.

2.1. PCA-CCA based domain adaptation

The classification is likely to perform poor if the distribution of training and test data does not match, which is always the case when two databases are recorded in different conditions. A PCA-CCA based domain adaptation method is applied to find the shared representation of features and diminish the negative consequences of cross-corpus training in the first stage of modelling. Two sets of principle components are learned from source (CreativeIT) and target (IEMOCAP) data and applied to both datasets to create two different views. This paired mapped data is then fed to linear CCA to obtain a shared view. The classifier is trained on M top dimensions with largest correlation from the mapped training data and tested on the mapped test data. The final dimensionality of frame-level feature vectors corresponds to the maximum number of CCA components kept M. The reader is referred to [8] for details.

2.2. First-stage LSTM modelling

Frame-level LSTM modelling considers only individual speaker's emotional behavior provided acoustical context. Under a reasonable assumption that emotions do not change rapidly from turn to turn, all the utterances within a dialog belonging to a single speaker are concatenated together and acoustical context is defined as a certain number of frames preceding the given frame within the stream. An example context window is depicted in Figure 2, where w stands for the context window length (in number of frames), F_t corresponds to the last frame of Turn N-1, F_0 is the first frame of Turn N, and both turns N-1 and N belong to the same speaker.

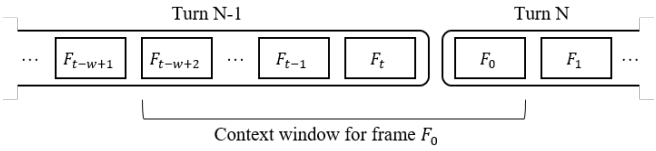

Fig. 2. An example of frame-level acoustical context window

Collected at the output of the first-stage LSTM are frame-level continuous arousal and valence predictions for individual speaker's emotional flow.

2.3. Balancing

Balancing refers to the preprocessing of the frame-level predictions of the first-stage LSTM to form the turn-change stream. The definition of a turn change is borrowed from [9]. Turn change consists of two turns. Each turn is defined as the portion of speech belonging to a single speaker (A or B) before he/she finishes speaking and may consist of multiple original segmented utterances. Consecutive utterances belonging to the same speaker are merged together as shown in Figure 3, where Turn A_1 is the first turn of speaker A, Turn B_1 is the first turn of speaker B and so on. The frame-level predictions are averaged to obtain a single value representation for each utterance. Thus, every turn change will contain a prediction for emotional primitive (activation or valence)

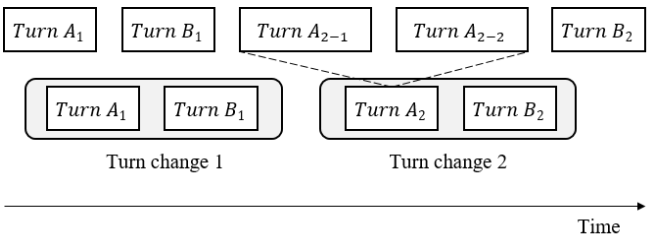

Fig. 3. Definition of a turn change. Two consecutive utterances Turn A_{2-1} and Turn A_{2-2} of speaker A are merged together to form a single turn A_2

for speaker A and the prediction of the same primitive for his partner B.

2.4. Second-stage LSTM modelling

Turn-level LSTM modelling considers the interaction between two speakers and predicts the emotional primitives for a given speaker provided the dialog history. The input to the LSTM network at every time step consists of a turn-change, i.e. the prediction of an emotional primitive for a particular utterance of speaker A, the prediction of the same emotional primitive for the utterance of his partner B within that same turn change, and additionally a gender tag for the speaker of interest. The predictions for every speaker are made independently: the history builds up on the utterances of the same speaker, and his partner's emotions are considered as an additional feature providing useful context.

3. EXPERIMENTS

3.1. Datasets

The first-stage LSTM is trained on the CreativeIT corpus and tested on the IEMOCAP corpus. Both datasets were recorded in the University of Southern California at 48kHz. The details are given below.

3.1.1. IEMOCAP

IEMOCAP corpus [6] contains acted emotional dialog conversations of 10 speakers (5 male and 5 female). The number of utterances in each dialog ranges from 10 to 90. Each utterance is on average 2-5 sec. long. The dialogs are recorded in English by experienced actors with prepared scenarios and improvised scripts from everyday life situations. In each dialog, there are 2 speakers with opposite gender. The total size of the corpus makes approximately 12 hours of audio-visual recordings. Only the audio modality is used in the current study. Audio recordings from both scripted and improvised dialogs are used for training and testing.

Original annotations for valence and activation in the IEMOCAP corpus are discrete values on a Likert scale in the range of 1 to 5. Every utterance is annotated with 2 or 3 annotators' scores plus the self-evaluation score, which are used to yield a single label by averaging all the scores together. Furthermore, the labels are clustered in three groups corresponding to low, medium and high range of values. The decision regions for activation are: [1, 2.5] low, (2.5, 3.5) medium, and [3.5, 5] high; for valence: [1, 2.3] negative, (2.3, 3.5) neutral, and [3.5, 5] positive. Such partitioning allows to balance the classes, and is in line with other clustering methods used in [9, 10]. The histograms of the distribution of valence and activation scores are depicted in Figures 4 and 5, respectively.

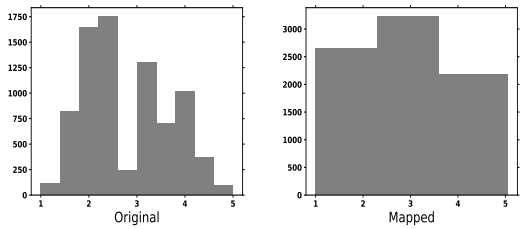

Fig. 4. Histogram of valence scores distribution

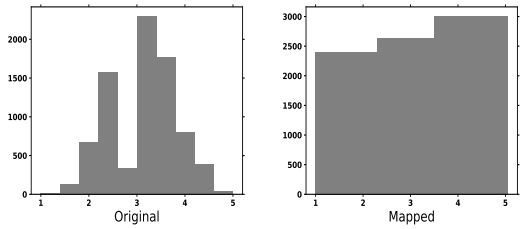

Fig. 5. Histogram of activation scores distribution

3.1.2. CreativeIT

CreativeIT corpus [7] contains acted emotional dialogs with pre-defined attitudes for each speaker. Only the dialogs recorded with the paraphrase technique that does not lexically restrict actors are used for training, which makes for a total of 31 recordings from 15 speakers with average duration of 4.3 minutes. The labels are provided by 3 raters with annotation frequency of 60 Hz.

3.2. Experimental setup

The classification task is to predict three levels of activation and valence for every utterance in IEMOCAP corpus given the dialog history. 130 Low Level Descriptors (LLDs) are extracted at 100 Hz using openSMILE toolkit [11] with a predefined configuration file from ComParE 2016 [12]. Cascaded normalization is applied as proposed in [13] prior to training. The PCA-CCA based domain adaptation is performed to find a new joint feature representation with the number of CCA and PCA components varying as 130, 100, 80 and 50, which corresponds to explained variance of 100%, 99.2%, 97.6%, and 91.4%. The first-stage LSTM model is trained on the mapped data from CreativeIT corpus following the approach adopted by Fedotov et al. [14], where the LSTM network consists of 2 hidden layers of 80 and 60 cells, respectively followed by dropout layers with probability of dropout p=0.3 to prevent overfitting. The optimizer is RMSprop with learning rate 0.001. Two models are learned for activation and valence separately using context corresponding to 3 seconds.

The architecture of the second-stage LSTM includes one hidden layer with 128 LSTM units followed by a fully connected layer used to output a single regression value, which is then thresholded according to decision regions described in

Section 3.1.1. The model is optimized by Mean Squared Error (MSE) loss with L2 regularization, beta=0.01 and Adam optimizer. A constant value of learning rate was set to 0.0001 in all the experiments. The maximum number of training epochs is set to 500 but early stopping is applied to prevent the network from overfitting. The number of turn changes in the dialog considered as context ranges from 5 to 20. The experiments were conducted separately for activation and valence with Leave-One-Session-Out cross-validation scheme and Unweighted Average Recall (UAR) performance measure. The performance of the proposed system is compared to a baseline that uses only frame-level context (i.e. first level LSTM) and ignores the dialog-level context as well as domain adaptation when making the utterance level predictions.

3.3. Results and Discussion

The summary of the results can be seen in Table 1. As expected, the proposed combination of using dialog context and domain adaptation achieves the best result of 76.3±1.9 for activation and 65.1±3.1 for valence. High deviation is an indicator of a large variance between the sessions. It is interesting to observe that the proposed method allows to significantly reduce the amount of variance compared to baseline systems.

The experiments showed that the optimal context length is 20 dialog turns. It is reasonable to assume that less amount of context does not capture enough history and loses modelling capability, while more context becomes confusing since older utterances have weaker relations to a given turn. The number of components in PCA-CCA based domain adaptation, which also has an effect of dimensionality reduction, significantly influences the performance of the system; the peak efficiency is observed around 50-80 components. Future work includes bridging the gap between activation and valence performance via multilingual approach, applying ASR systems and sentiment analysis techniques for extracting complimentary lexical features [15].

3.4. Relation to prior work

An inspiration for cross-corpus and cross-task LSTM setting was found in the recent work by Kaya et al. [16]. The application of domain adaptation technique, firstly proposed by Sagha [8] on the utterance level, is exploited on LLD level for the first time in this work. Importance of domain adaptation was also shown in [17]. Several works dedicated to hierarchical contextual emotion modelling became the basis for comparing the efficiency of the proposed method. Metallinou et al. [10] employed a two-level classification scheme with Hidden Markov Model (HMM) using audio features, and Bidirectional LSTM (BLSTM) using both audio and visual features. They report Unweighted Accuracy obtained in cyclic Leave-One-Speaker-Out cross-validation experiments. Lee et al. [9] used speech based features with Dynamic Bayesian Network (DBN) structure to model time and cross-speaker dependencies between two interacting partners' emotional states. They report accuracy percentage in a 15-fold cross-validation experiment. The proposed method outperforms both approaches, as can be seen in Tables 2 and 3, respectively.

Table 1. Classification results on IEMOCAP corpus (UAR, %) in comparison to baseline systems. +Con: with context, +Ada: with PCA-CCA based adaptation

	Baseline	+Ada	+Con	+Ada +Con
Aro	46.5±7.3	57.8±7.0	68.2±8.9	**76.3**±1.9
Val	50.0±8.2	56.4±5.9	57.3±9.5	**65.1**±3.1

Table 2. Classification results on IEMOCAP (UAR, %) in comparison to Metallinou et al. [10], speaker-independent cross-validation, audio (a) and visual (v) features

	[10] (a)	[10] (a+v)	Proposed (a)
Arousal	61.9±4.9	52.3±5.4	**76.3**±1.9
Valence	50.0±3.6	**64.7**±6.5	**65.1**±3.1

Table 3. Classification results on IEMOCAP in comparison to Lee et al. [9], 15-fold cross-validation, audio features only

	[9] (Acc)	Proposed (UAR)
Arousal	63.5	**77.7**
Valence	65.0	**69.4**

4. CONCLUSION

This paper proposes a novel hierarchical two-level LSTM based model for classifying valence and arousal emotional primitives into three categories: low, medium and high. Proposed approach provides robust predictions taking into account both individual speaker emotional behavior as well as interaction between two speakers and the dialog level temporal dynamics. Cross-corpus framework for lower level LSTM modelling allows to train on more data, and handles the lack of continuous annotation in the target corpus. A domain adaptation method is newly employed at LLD level to cope with the distribution mismatch. The proposed method proves generalizable and effective by a set of Leave-One-Session-Out cross-validation experiments that reach UAR scores of 76.3% and 65.1% for activation and valence, respectively. This outperforms previous results shown in literature with higher mean and lower standard deviation across the folds, which advances the state-of-the-art using only audio modality.

Acknowledgements. The study is supported by the Russian Science Foundation (project No. 18-11-00145) and Huawei Innovation Research Program.

5. REFERENCES

[1] Ashish Tawari and Mohan Manubhai Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Transactions on multimedia*, vol. 12, no. 6, pp. 502–509, 2010.

[2] Jackson Liscombe, Giuseppe Riccardi, and Dilek Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[3] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[4] Wootaek Lim, Daeyoung Jang, and Taejin Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–4.

[5] Zixing Zhang, Fabien Ringeval, Jing Han, Jun Deng, Erik Marchi, and Björn Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with lstm neural networks," in *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, 2016, pp. 3593–3597.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[7] Angeliki Metallinou, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

[8] Hesam Sagha, Jun Deng, Maryna Gavryukova, Jing Han, and Björn Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*. IEEE, 2016, pp. 5800–5804.

[9] Chi-Chun Lee, Carlos Busso, Sungbok Lee, and Shrikanth S Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[10] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[11] Florian Eyben, Felix Weninger, Florian Groß, and Björn Schuller, "Recent Developments in openSMILE, the Munich open-source Multimedia Feature Extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 2013, pp. 835–838.

[12] Björn W Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron C Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *INTERSPEECH, Proceedings*, 2016, pp. 2001–2005.

[13] Heysem Kaya and Alexey A Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, 2018.

[14] Dmitrii Fedotov, Denis Ivanko, Maxim Sidorov, and Wolfgang Minker, "Contextual dependencies in time-continuous multidimensional affect recognition," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 1220–1224.

[15] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[16] Heysem Kaya, Dmitrii Fedotov, Ali Yeşilkanat, Oxana Verkholyak, Yang Zhang, and Alexey Karpov, "LSTM based cross-corpus and cross-task acoustic emotion recognition," in *Proc. INTERSPEECH 2018*, 2018, pp. 521–525.

[17] Mohammed Abdelwahab and Carlos Busso, "Supervised domain adaptation for emotion recognition from speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5058–5062.