Contextual Dependencies in Time-Continuous Multidimensional Affect Recognition

Dmitrii Fedotov [1,2], **Denis Ivanko** [1,2], **Maxim Sidorov** [1], **Wolfgang Minker** [1]
[1] Ulm University, [2] ITMO University
Ulm Germany, St. Petersburg Russia
{dmitrii.fedotov, wolfgang.minker}@uni-ulm.de,
denis.ivanko11@gmail.com, maxim.sidorov@alumni.uni-ulm.de

Abstract

Modern research on emotion recognition often deals with time-continuously labelled spontaneous interactions. Such data is much closer to real world problems in contrast to utterance-level categorical labelling in acted emotion corpora that have widely been used to date. While working with time-continuous labelling, one usually uses context-aware models, such as recurrent neural networks. The amount of context needed to show the best performance should be defined in this case. Despite of the research done in this field there is still no agreement on this issue. In this paper we model different amounts of contextual input data by varying two parameters: sparsing coefficient and time window size. A series of experiments conducted with different modalities and emotional labels on the RECOLA corpora has shown a strong pattern between the amount of context used in model and performance. The pattern remains the same for different pairs of modalities and label dimensions, but the intensity differs. Knowledge about an appropriate context can significantly reduce the complexity of the model and increase its flexibility.

Keywords: time-continuous affect recognition, affective context analysis, multimodal emotion recognition.

1. Introduction

Real life human-human interaction consists of two main aspects: information contained in speech and emotion expressed by humans. Speech recognition techniques allow computers to understand human speech but they lack an emotional component. Exactly the same words or phrases may be a statement, a question or a guess if said with different emotions. These types of phrases should be recognised and processed by dialogue system differently. It is necessary for computers to understand human emotions in order to succeed in interaction with them.

Modern research focus on natural interaction between computer-based systems and humans. These systems try to understand non-standardized questions and provide answers, similar to the spontaneous interaction between two humans. One of the most important parts of human understanding is the ability to identify and react to emotions. Emotion recognition may significantly improve the quality of human-computer interaction, speech recognition systems and artificial intelligence in general.

Previous research on emotion recognition mostly dealt with utterance-level categorical data labelling, i.e. each data sample had one label from the list, e.g. anger, happiness, neutral, etc. However, recent research has focused on the dimensional time-continuous data that provides more flexibility and precision of emotion definition. This type of data requires more complex models and the definition of additional parameters, such as the amount of context to be used. Despite of the research conducted in this area, it is still an open question, how much previous data do the system need to provide the best performance. Studies on the effect of this parameter will help to build an effective end-to-end real time emotion recognition system.

As shown in this paper, there is a strong correlation between the amount of context and the performance of an emotion recognition system despite of the amount of data i.e. the time steps used.

This paper is structured as follows: Section 2 provides an overview of research related to multi-dimensional time-continuous emotion recognition; Section 3 details the data used in this study as well as the pre- and postprocessing procedures; in Section 4 the methodology used is described; in Section 5 experimental results are shown and analysed; conclusions from this study and proposed future research are presented in Section 6, followed by acknowledgements in Section 7.

2. Related work

Previous research on emotion recognition mostly dealt with utterance-level categorically labelled databases. (Haq and Jackson, 2010; Burkhardt et al., 2005; Makarova and Petrushin, 2002). Corpora with time-continuous labelling emerged in the past years and they gain popularity among researchers (Schroeder et al., 2012; Ringeval et al., 2013). Time-continuous emotion recognition provides more flexibility for the system, but also creates new challenges. Firstly, is the amount of previous information that should be used to model emotions. According to Levenson, 1988 it should be a value between 0.5 and 4 seconds, but is still remains an open question and depends on modality and emotional dimension (Gunes and Pantic, 2010).

Another issue refers to the labelling process of emotional interactions. Annotations of emotions are performed by humans, hence, they yield a significant level of subjectivity and a suitable method is required for computing a gold standard. It can be based on correlation between individual ratings provided by annotators (Mariooryad and Busso, 2013; Nicolle et al., 2012).

When annotating time-continuous emotions, a reaction lag may also appear; therefore, it should be considered when synchronising features and labels. It may be done by maximising the correlation between some features and emotional ratings and/or ratings from annotators (Nicolaou et al., 2010; Mariooryad and Busso, 2014).

Another related issue is the appropriate sampling frequency of data. Original data can be upsampled or downsampled to thr required value of frequency (Nicolaou et al., 2011; Metallinou et al., 2011). In this paper we used data sparsing to study the effect of amount of data combined with its intensity on system performance.

3. Data and data-related procedures

The recently introduced time-continuously labelled corpus of spontaneous interaction in French called RECOLA (Remote COLlaborative and Affective interactions) (Ringeval et al., 2013) was used in this paper.

3.1 Corpus description

The RECOLA database was collected during the resolving of a cooperative problem. It consists of spontaneous interactions between 23 dyadic pairs of French-speaking participants, i.e. 46 persons. 34 participants gave their consent to share the data. The dataset was therefore reduced from originally recorded 9.5 to 7 hours. Annotations for 23 of them are publically available in current version of database and were used in this research.

The participants were aged between 18 and 25 years and have different mother tongue although they spoke French during recorded interactions: 17 of them had French as a mother tongue, 3 – Italian and 3 – German.

RECOLA consists of recordings in 4 modalities: audio, video, electro-cardiogram and electro-dermal activity. Interactions were evaluated by 6 equally gender distributed French speaking annotators through ANNEMO (ANNotating EMOtions) tool (Ringeval et al., 2013). The annotations include two emotional (arousal and valence) and five social (agreement, dominance, engagement, performance and rapport) behaviour dimensions.

3.2 Features

Audio features were extracted with the openSMILE open-source software. The feature set consists of 3 groups of low-level descriptors (LLDs): prosodic, spectral, cepstral and voice quality. (Eyben et al., 2010; Schuller et al., 2014). There are 65 LLDs and along with their first order derivate we have 130 features in total.

The visual feature set consists of 20 LLDs and their first order derivate for each frame available. It includes 15 facial action units, 3-dimansional head pose and the mean and standard deviation of the optical flow in the region around the head (Ringeval et al., 2013).

3.3 Data preprocessing

The available part of the dataset was divided into 2 speaker disjoint subsets: train and evaluation. Subsets maintain age, gender and mother tongue distribution of original set (see Table 1).

Set	Age μ(σ)	Gender	Mother tongue
Full	21.35 (2.04)	10 males 13 females	17 French 3 Italian 3 German
Train	21.38 (2.13)	7 males 9 females	12 French 2 Italian 2 German
Evaluation	21.29 (1.98)	3 males 4 females	5 French 1 Italian 1 German

Table 1: Partitioning of RECOLA database into train and evaluation subsets

3.3.1 Features and labels preprocessing

Provided audio and video features were normalised with the Z-transformation based on the train subset.

There are two ways of using several ratings for each recording: merge them into a gold-standard rating or train model to produce separate predictions. The first approach was used in this paper. The gold-standard may be calculated by simple averaging the values provided by each evaluator. However, this methodology may lead to a loss of information contained in annotator's perception of emotions. It effects the spread of emotional ratings as well as their "neutral" values, i.e. bias. For example, some annotators perceive emotions in a mild manner (e.g. Annotator 6) rather than in a strong one (Annotator 1). At Figure 1 a diversity of ratings from annotators of RECOLA database for all recordings available is shown.

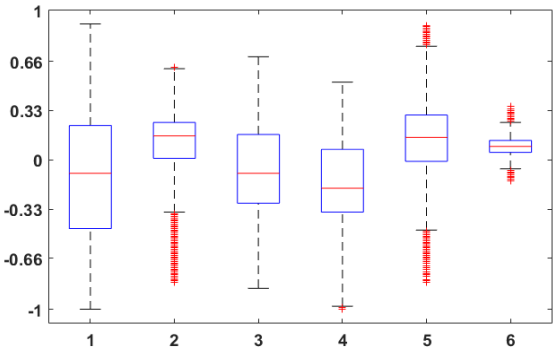

Figure 1: Ratings diversity from 6 annotators of RECOLA database

Taking these differences into account, the gold-standard was based on the maximisation of the inner-rater agreement (Mencattini et al., 2016).

While estimating emotions continuously, evaluators need some time to report the changes. The delay between an actual change of emotional behaviour and the moment it is annotated is called reaction lag (RL). It is not consistent for different speakers and label dimensions, although has negligible variation for different annotators (Mariooryad and Busso, 2014).

The value of RL may be calculated based on the correlation of features with labels. Some features are strongly correlated with positive values of particular labels, some with negative. In previous research the *RL* was found to be 3.89 s for arousal and 4.52 s for valence (Mencattini et al., 2016; Ringeval et al., 2015; Mariooryad and Busso, 2014). The value of RL for arousal was corrected to be 3.88 due to label rate (25 Hz).

Gold-standard labels were shifted backwards according to RL values mentioned above. The label values for the last frames of each speaker were lost after shifting and replaced with zeros. Labels were normalised with Z-transformation based on train subset and denormalised at estimation stage.

3.3.2 Contextual pre- and postprocessing and sparsing

To meet the requirements of the context-based model, features and labels were preprocessed from [samples × features(labels)] representation to [samples × time steps × features(labels)]. Time steps were taken only backwards for both features and labels. The time window size (TW) defines the number of previous steps to be taken for every sample in set. Previous frames could be used only if they exist and relate to the same speaker as the current frame, otherwise zero-padding was applied.

This procedure was combined with sparsing. If the sparsing coefficient (SC) is greater than one, then every *n-th* frame is taken into account at the stage of adding time steps. For example, for sample *t* with *SC=3* and *TW=6* the following frames are chosen at contextual preprocessing stage: *[t-15, t-12, t-9, t-6, t-3, t]*.

The combination of *TW* and *SC* define the amount of context (in seconds) used by the model:

$$C = \frac{SC \times TW}{frame\ rate}$$

The same context may be represented with different pairs of *SC* and *TW*. For example, the context is equal to 24 frames with [*SC=3, TW=8*] and [*SC=1, TW= 24*]. To make a uniform grid, the following values of contextual parameters were chosen: *TW = {2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64}, SC = {1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64}.*

A sequence-to-sequence approach was applied in this research, i.e. features of *TW* previous frames were used to make a prediction of labels for the same *TW* previous frames. When the predictions were made, values obtained for the same frame at different time steps were average to smooth the final prediction.

4. Methodology

Traditional feed-forward neural networks are not capable of using the previous information as they lack feedback connections in their architecture. This leads to constraining the knowledge about the data only to the current frame. To overcome this problem, recurrent neural networks (RNN) with one time step of delay were introduced. However, networks of this type suffer from the vanishing gradient problem and cannot store the information about more than approximately 10 time steps (Hochreiter et al., 2001). To avoid the problem of exponentially decaying gradients, a long-short term memory recurrent neural network (LSTM) was introduced (Hochreiter and Schmidhuber, 1997) and then improved (Graves and Schmidhuber, 2005). LSTM-RNNs use a fine regulation of the system state by special gates: input gate, output gate and forget gate. These gates allow to accumulate the information about previous time steps over long duration and drop the information when needed. The weights of self-loops are not fixed, but based on the gates which allows to change the level of data integration.

Two LSTM layers with 20 and 15 neurons respectively were used in this research. The LSTM blocks had a ReLU activation function (Vinod and Hinton, 2010) and the neurons of the output layer had a simple linear activation function. To avoid overfitting, recurrent layers were followed by the dropout layers (Srivastava et al., 2014). Different dropout probability values were studied and *p=0.1* was selected as it provided the best results. The LSTM models were optimized by root mean square propagation (RMSprop) using the concordance correlation coefficient (CCC) as a metric function. LSTM implementation is provided by Keras (Chollet, 2015).

Our previous research has shown, that the performance of the system based on RNN significantly depends on the learning rate of optimiser. While too high values of the learning rate cannot provide any appropriate performance (zero performance), rather small ones may result in slow learning and the system may permanently get stuck. The value of learning rate, that provides the lowest loss is usually the highest one before the "zero performance" values, (see Figure 2).

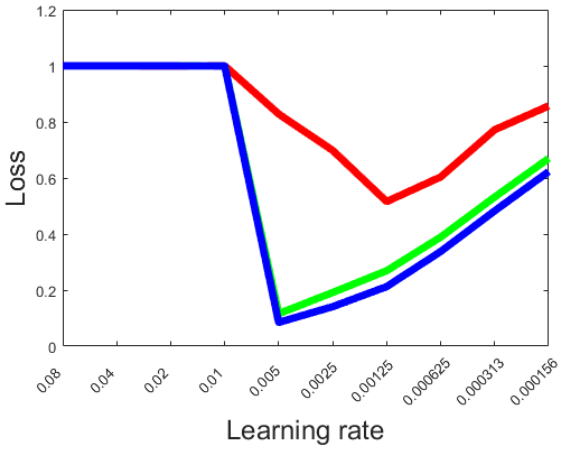

Figure 2: Learning rate against loss.
Red – after 1st epoch; green – after 20th epoch; blue – after 100th epoch.

The appropriate learning rate varies with the number of previous time steps used; therefore, it cannot be fixed. As it has a critical effect on performance, an automatic procedure of the learning rate selection was developed. According to LeCun et al. (1998), the search of the best learning rate was conducted with the decreasing factor of 2. The following values were used: 0.08, 0.04, 0.02, 0.01, 0.005, 0.0025, 0.00125, 6.25e-4, 3.13e-4, 1.56e-4. The model training was started with the highest learning rate. The value of a loss function was calculated on the train sample after the first epoch. If the loss was greater than a predefined threshold (empirically set to *t=0.85*), training was terminated and the next learning rate was tried out. The training loss was further monitored during the training process and if the loss at current epoch was greater than at the first one, the same procedure of training termination was applied.

5. Experiments and results

A series of experiments was conducted to study the impact of context on the performance of an emotion recognition system. All pairs of contextual parameters described above were used for audio and video modalities as well as their feature-based multimodal fusion for two emotional dimensions.

The performance of the emotion recognition system on evaluation subset was estimated with CCC. The results are shown at Figure 3. The results were obtained for each pair of sparsing coefficient and time window size and interpolated afterwards to create a surface of system performance, indicated with colour map. Diagonal lines represent the amount of used context in seconds. Red stars show the best sparsing coefficient that led to the best performance of system with each value of *TW*. The red stars in a circle represent the best performance obtained within the provided problem definition. One may notice a strong pattern between the amount of context and performance of the emotion recognition system. It is especially obvious with Audio-Arousal pair. Patterns may still be noticed in Audio-Valence and Video-Valence pairs.

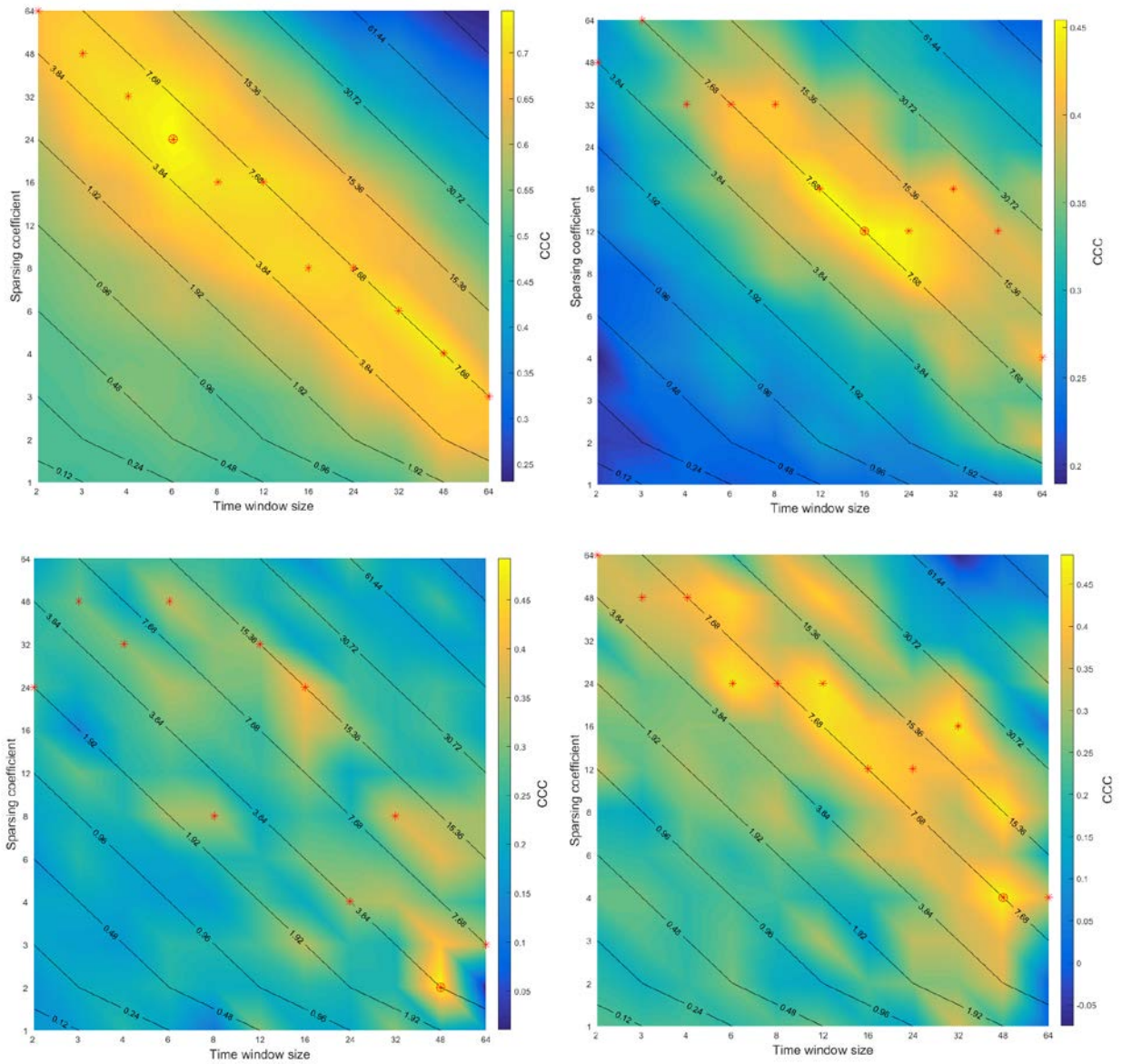

Figure 3. Contextual dependencies of performance in RECOLA database.
Top-left: Audio-Aruosal; Top-right: Audio-Valence; Bottom-left: Video-Arousal; Bottom-right: Video-Valence.

Results for Video-Arousal do not show any trustworthy trends and the emotion recognition system does not perform well with this modality-label pair. The same methodology but with 80 and 60 neurons respectively was tried and it showed the same trends.

Patterns for different modality-label pairs are similar and the best results are lying in the area of approximately 6 seconds of context for arousal and 8 seconds for valence. Performance of the system obtained with different combinations of contextual parameters does not differ much; therefore, less amount of data may be used to obtain the same high results.

6. Conclusion

Experiments have shown a strong pattern between the amount of context and the performance of an emotion recognition system. Sparsing does not affect performance much, while allowing to use more simple and flexible models and get results much faster. The knowledge about sparsing coefficients may reduce the number of time steps for RNNs to 6-12. The information about the appropriate amount of required contextual data may be used in real-time emotion recognition systems.

Further research will be focused on other time-continuously labelled corpora, such as SEMAINE and context-aware models (e.g. Hidden Markov Models and Gated Recurrent Units).

7. Acknowledgements

The work presented in this paper was partly supported by the DAAD (German Academic Exchange Service) within the different programmes and by the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” which is funded by the German Research Foundation (DFG).

8. Bibliographical References

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional

speech. In Ninth European Conference on Speech Communication and Technology.
Chollet F. Keras. (2015). URL: https://github.com/fchollet/keras.
Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, pp. 1459-1462.
Gunes, H., Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. Inter. Journal of Synthetic Emotions 1, pp. 68–90.
Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6), pp. 602-610.
Haq, S. and Jackson, P., (2010). Machine Audition: Principles, Algorithms and Systems, chapter Multimodal Emotion Recognition, IGI Global, Hershey, pp. 398-423.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), pp. 1735-1780.
Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.
LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp. 2278-2324.
Levenson, R. W. (1988). Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. Social psychophysiology: Theory and clinical applications., pp. 17–42.
Makarova, V. and Petrushin, V. (2002). Ruslana: A database of russian emotional utterances. In Proc. Int. Conf. Spoken Language Processing.
Mariooryad, S., Busso, C. (2013). Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations, in: Proc. of the 5th Inter. Conf. on Affective Computing and Intelligent Interactions (ACII), Geneva, Switzerland. pp. 85–90.
Mariooryad, S., & Busso, C. (2015). Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. IEEE Transactions on Affective Computing, 6(2), pp. 97-108.
Mencattini, A., Martinelli, E., Ringeval, F., Schuller, B., & Di Natale, C. (2017). Continuous estimation of emotions in speech by dynamic cooperative speaker models. IEEE transactions on affective computing, 8(3), pp. 314-327.
Metallinou, A., Katsamanis, A., Wang, Y., & Narayanan, S. (2011). Tracking changes in continuous emotion states using body language and prosodic cues. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 2288-2291
Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning, pp. 807-814.
Nicolaou, M., Gunes, H., Pantic, M. (2010). Automatic segmentation of spontaneous data using dimensional labels from multiple coders, in: Proc. of the LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing, Istambul, Turkey. pp. 43-48.
Nicolaou, M. A., Gunes, H., & Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Transactions on Affective Computing, 2(2), pp. 92-105.
Nicolle, J., Rapp, V., Bailly, K., Prevost, L., & Chetouani, M. (2012). Robust continuous prediction of human emotions using multiscale dynamic cues. In Proceedings of the 14th ACM international conference on Multimodal interaction, pp. 501-508.
Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J. P., Ebrahimi, T. & Schuller, B. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognition Letters, 66, pp. 22-30.
Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D. & Pelachaud, C. (2012). Building autonomous sensitive artificial listeners. IEEE Transactions on Affective Computing, 3(2), pp. 165-183.
Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y. (2014). The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load, in: Proc. of INTERSPEECH 2014, 15th Annual Conf. of the Inter. Speech Communication Association (ISCA).
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), pp. 1929-1958.