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

A Saliency-based Attention LSTM Model for Cognitive Load Classification from Speech

Ascensión Gallardo-Antolín[1], *Juan M. Montero*[2]

[1]Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid, Spain
[2]Speech Technology Group, ETSIT, Universidad Politécnica de Madrid, Madrid, Spain

gallardo@tsc.uc3m.es, juancho@die.upm.es

Abstract

Cognitive Load (CL) refers to the amount of mental demand that a given task imposes on an individual's cognitive system and it can affect his/her productivity in very high load situations. In this paper, we propose an automatic system capable of classifying the CL level of a speaker by analyzing his/her voice. Our research on this topic goes into two main directions. In the first one, we focus on the use of Long Short-Term Memory (LSTM) networks with different weighted pooling strategies for CL level classification. In the second contribution, for overcoming the need of a large amount of training data, we propose a novel attention mechanism that uses the Kalinli's auditory saliency model. Experiments show that our proposal outperforms significantly both, a baseline system based on Support Vector Machines (SVM) and a LSTM-based system with logistic regression attention model.

Index Terms: cognitive load, speech, LSTM, weigthed pooling, auditory saliency, attention model

1. Introduction

Cognitive Load (CL) refers to the amount of mental demand that a given task imposes on a subject's cognitive system and it is usually associated to the working memory that refers to the capacity of holding short-term information in the brain [1]. As overload situations can affect negatively the individual's performance, the automatic detection of the cognitive load levels has many applications in real scenarios such as drivers' monitoring.

Speech-based CL detection systems are particularly interesting since they are non-intrusive and speech can be easily recorded in real applications. In fact, in 2014, an international challenge (Cognitive Load Sub-Challenge inside the INTERSPEECH 2014 Computational Paralinguistics Challenge) was organized with the aim of studying the best acoustic features and classifiers for this task [2]. Following this line of research, this work focuses on the design of an automatic system for CL level classification from speech.

Regarding the feature extraction, different acoustic characteristics have been proposed as spectral-related parameters such as, Mel-Frequency Cepstral Coefficients (MFCC) [2], [3], spectral centroid, spectral flux [2] and prosodic cues (intensity, pitch, silence duration, ...) [4], [5]. Respect to the classifier module, Gaussian Mixture Models (GMM) [3] and Support Vector Machines (SVM) [2] [5] are the most common choices.

However, in the last years, the application of Deep Learning (DL) models to speech-related tasks, such as Automatic Speech Recognition (ASR) [6], [7], Language Recognition (LR) [8] or Speech Emotion Recognition (SER) [9], [10], [11] has allowed to increase the performance drastically. Convolutional Neural Networks (CNN) [12], [7], Long Short-Term Memory (LSTM) [13] and their combination are the most commonly used architectures in this field. On the one hand, CNNs exhibit the capability of learning optimal speech representations. On the other hand, LSTMs are capable to perform temporal modeling, so they are very suitable for dealing with sequences as it is the case of speech signals.

A new line of research, complementary to CNN and LSTM models, tries to learn the structure of the temporal sequences aiming at modeling the relevance of each frame to the task under consideration. In particular, the so-called attention models are capable of increasing the DL-based systems performance by emphasizing the contribution of certain temporal frames to the final output. These models have been successfully proposed for ASR [14], machine translation [15] or SER [9], [10], [11], [16].

From a more general perspective, the concept of attention refers to a complex cognitive function that allows humans to select the most salient or relevant events in their environment in order to focus their sensory and cognitive resources on them. Regarding the aural modality, in recent years, there have been several efforts for developing attention or saliency auditory models that try to mimic this human mechanism [17], [18]. In this paper, we hypothesize that these saliency models can help to determine the frames conveying the most relevant information about the subject's CL level, and, for that, they can be used as a kind of attentional model inside a LSTM-based system.

In this paper, we present two main contributions. Firstly, we focus on the use of LSTMs in combination with different weighted pooling strategies for CL level classification, as, to our knowledge, there is no previous works in this direction for this specific task. As this problem has many similarities to SER, our work is mainly based on previous research on emotion classification from speech, especially, on [9] and [11]. Secondly, we propose the use of auditory saliency models as an attentional mechanism for LSTM-based CL level determination.

The remainder of this paper is organized as follows: Section 2 describes the fundamentals of LSTM with weighted pooling, Section 3 covers the attention-based weighting schemes considered, including our proposal. Results are presented in Section 4, followed by some conclusions of the research in Section 5.

2. LSTM with Weighted Pooling

Long Short-Term Memory networks are a kind of recurrent neural networks that have the ability to store information from the past in the so-called memory blocks [13], in such a way that they are capable of learning long-term dependencies, overcoming the vanishing gradient problem. Therefore, LSTM outputs depend on the present and previous inputs, and, therefore, they are very suitable for modeling temporal sequences, as speech.

The sequence-to-sequence learning carried out by LSTMs can be thought as a transformation of an input sequence of

http://dx.doi.org/10.21437/Interspeech.2019-1603

length T, $x = \{x_1, ..., x_T\}$ into an output sequence $y = \{y_1, ..., y_T\}$ of the same length, assuming that the classification process is easier in the y-space than in the x-space. However, as in the case of SER, CL classification can be seen as a many-to-one sequence-to-sequence learning problem [9], as the input is a sequence of acoustic vectors and the final system output must be the predicted cognitive load level for the whole utterance (one single value). For this reason, it is advisable to include an intermediate stage in order to generate a more compact representation of the temporal LSTM output sequence that, in turn, will be the input to the classifier [9], [10]. A most common option is to use a Weighted Pooling (WP) module [11] consisting of two different steps, weighting and temporal integration.

In the first stage, a weight α_t is computed and assigned to each temporal LSTM output y_t, following a certain criterion. In the second one, temporal aggregation, the weighted elements of the LSTM output sequence are somehow combined over time for producing a summarized representation of the information contained in this LSTM output. Usually, this is done by performing a simple aggregation operation, as follows,

$$z = \sum_{t=1}^{T} \alpha_t y_t \tag{1}$$

where $y = \{y_1, y_2, ..., y_T\}$ is the LSTM output sequence, $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_T\}$ is the corresponding weight vector and z is the final utterance-level representation.

Well-known particular examples of WP are *mean-pooling*, where all the weights are equal and set to $1/T$ [11]; *max-pooling*, where all weights are zero except the weight of the maximum observed output, which is 1 [12]; and *last-pooling* where only the weights corresponding to the last M frames of the LSTM output are different from zero and set to $1/M$ [8].

3. Attention-based Weighted Pooling Schemes

In this Section we present the two attentional WP schemes used in this work: the regression attention weights described in [11] for SER and our proposal based on auditory saliency models.

3.1. Logistic regression attention weights

According to [11], this strategy is appropriate in situations where there is a lack of training data (as in this case), preventing the use of more complex attention models, as those described in [9], [10]. Here, the weights are computed as a simple logistic regression, through this equation,

$$\alpha_t = \frac{\exp(u^T y_t)}{\sum_{t=1}^{t=T} \exp(u^T y_t)} \tag{2}$$

where u and y are the attention parameters and the LSTM output, respectively. Both, u and y are obtained by using the back-propagation algorithm in the system training process.

3.2. Saliency-based weights

Our hypothesis is that, when the training data is scarce, it is not feasible to properly trained attentional models, and therefore, it could be more effective to use attention weights derived from external cues. In particular, in this work, as external source of information, we consider the auditory saliency model developed by Kalinli [17] that assigns a saliency score to each time instant. Our assumption is that frames with higher saliency values are more likely to present a strong content about the subject's CL level, and therefore, larger weights should be assigned to them into the WP scheme.

Kalinli's model extracts five features (intensity, frequency and temporal contrast, orientation and pitch) from the spectrogram at multiple scales, that, after the computation of center-surround differences, result in a set of conspicuity maps. These maps are normalized by using an iterative and non-linear algorithm that emphasizes the most prominent areas in the time-frequency representation, and summed for obtaining the final 2D auditory saliency map.

In order to determine the relevance score for each temporal frame, the saliency map is summed across frequency channels for each time instant [19] and normalized to zero mean and unit variance at utterance-level, yielding a saliency signal $x_{sal}(t)$. Finally, the weights are obtained as the result of the softmax transformation applied to the saliency signal in order to guarantee that their sum across all the frames of the utterance is one,

$$\alpha_t = \frac{\exp(x_{sal}(t))}{\sum_{t=1}^{t=T} \exp(x_{sal}(t))} \tag{3}$$

We have considered two variants of this method. In the first case (*Normalized Saliency*), the saliency signal is derived from the normalized conspicuity maps, as explained before. In the second case (*Unormalized Saliency*), the only difference is that $x_{sal}(t)$ is obtained from the conspicuity maps without the application of the iterative normalization process.

4. Experiments and Results

4.1. Database and Baseline System

We have adopted the "Cognitive Load with Speech and EGG" (CSLE) database [20], [2] for our experiments. This database have been used in the Cognitive Load Sub-Challenge inside the 2014 COMPARE Challenge [2]. It contains speech from 26 Australian English speakers recorded while performing a set of tasks designed for inducing different levels of cognitive load (low, medium and high, denoted as *L1*, *L2* and *L3*, respectively). As in the challenge, we have considered these three tasks:

- *Reading Sentence (RS)*. Speakers were asked to read a set of sentences and recall an isolated letter between them.
- *Stroop Time Pressure (STP)*. Based on the Stroop test [21], speakers were required to indicate the color of a set of printed words that, in turn, are names of colors.
- *Stroop Dual (SD)*. Similar to the previous task, with the difference that speakers had to execute another simultaneous task (tone counting) in the high load scenario.

The challenge organizers provided a partition of the database into training + development and test subsets, where it was guaranteed that speakers belong to only one of these subsets. Table 1 shows the details about the database composition.

The baseline system is the one provided by the challenge organizers whose details can be found in [2]. In summary, the acoustic features are obtained by using the open-source openSMILE feature extractor [22] and are composed of 6373 static characteristics, which are functionals (statistical moments, percentiles, peaks, etc.) of short-term Low-Level Descriptors (LLD) and are computed at utterance level. The LLDs consist of 65 energy-related, spectral and prosodic characteristics computed in a short-term basis. The classifier is a linear kernel SVM implemented by using the WEKA toolkit [23].

Table 1: *Composition of the CSLE database. For each task, the number of utterances per subset and CL level are indicated.*

Task	Number of utterances				
	Subset		L1	L2	L3
Reading Sentence	Train+Dev	1350	378	378	594
	Test	600	168	168	264
Stroop Time Pressure	Train+Dev	162	54	54	54
	Test	72	24	24	24
Stroop Dual	Train+Dev	162	54	54	54
	Test	72	24	24	24
Total	Train+Dev	1674	486	486	702
	Test	744	216	216	312

4.2. LSTM-based Systems Configuration

Fig. 1 shows the LSTM architectures used in this work with two different weighting schemes: logistic regression weights (a) and our proposal, saliency-based weights (b). Both systems were implemented with the Tensorflow [24] and Keras [25] packages.

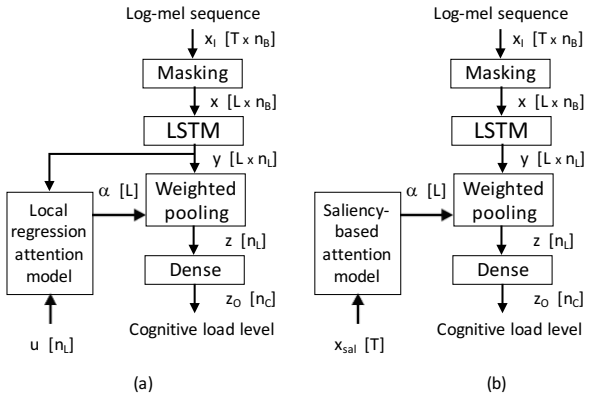

Figure 1: *Block diagram of the LSTM-based systems for CL classification: (a) Logistic regression model; (b) Saliency-based model. In brackets, the dimension of each variable, where T, L, n_B, n_L and n_C, stand for the no. of frames of the input signal, the length of the LSTM input/output sequence, the no. of mel filters, LSTM units and classes (CL levels), respectively.*

In all cases, the input feature set consists of $n_B = 64$ log-Mel filterbank energies (log-Mels) computed every 10 ms over Hamming windows of 32 ms long and a mel-scaled filterbank composed of 64 filters and is obtained by using the Librosa Python toolkit [26]. After feature extraction, mean and standard deviation normalization are applied at utterance-level yielding to a set of normalized log-Mels sequences x_I with T x n_B dimensions, where T is the number of frames of each utterance.

The length of the LSTM input sequences is set to $L = 1024$ (which corresponds to approximately 10 s) for the *RS* task and to $L = 2048$ (about 20 s) for the *STP* and *SD* tasks. Shorter utterances are padded with zeros by using a Masking layer, in such a way that these masked values are not used in further computations. Longer utterances are cut (this is only necessary in a few cases in the *SD* task). The output sequence of the Masking layer is denoted as x and its dimensions are L x n_B.

This sequence is passed through a LSTM layer with $n_L = 128$ memory cells and 25% dropout to avoid over-fitting in the training process. The LSTM output, y, is a sequence of size L x n_L. Next, the information contained in y is summarized by using the considered weighting scheme with weights α, yielding a n_L-dimensional vector, z. The length of the weight vector α is L. Note that when $T < L$, $\alpha_t = 0,\ T < t \leq L$. The vector z is the input of a dense layer with $n_C = 3$ nodes with softmax activation producing a n_C-dimensional output, z_O, representing the probabilities of each class (CL levels). Finally, the class with higher probability is assigned to the utterance.

The LSTM models are trained using stochastic gradient descent and the Adam method with an initial learning rate of 0.001 for the *RS* task and 0.0005 for the *STP* and *SD*. Following the challenge recommendations, each task is considered separately.

In the logistic regression model, the attention parameter vector u has a dimension of $n_L = 128$, all its components are initialized to $1/n_L$ and then refined during the training stage.

In our approach, T-dimensional saliency signals, denoted as x_{sal}, are derived from the spectrogram magnitude computed over Hamming windows of 37 ms long with a 95% overlap, following the Kalinli's auditory saliency model [17] as implemented in the MT_TOOLS toolbox [27].

4.3. Results

Table 2 contains the results achieved for the baseline system and different LSTM architectures for the three tasks under consideration, *RS*, *STP* and *SD*. The column "Average" refers to the micro-average across the tasks. As the number of instances for each class (CL levels) is unbalanced, results are given in terms of the Unweighted Average Recall (UAR) that is computed as the unweighted mean of the class-specific recalls. In the case of the LSTM-based systems, each experiment was run 10 times and therefore, Table 2 shows the average UAR across the 10 subexperiments and the respective standard-deviation.

LSTM corresponds to the conventional approach where no weighted pooling is applied and only the last frame of the LSTM output is passed through the following dense softmax layer. In the *LSTM+VAD* alternative, a Voice Activity Detector (VAD) is applied to the raw speech signals before the feature extraction process in order to remove the silence/noise frames. As can be observed, the use of a VAD is not beneficial as it produces a decrease in performance. This suggests that silence pauses convey important information for discriminating between different CL levels, as they are related to the rhythm, elocution speed and disfluencies that can be heavily affected by the speaker's cognitive load state [28], [29].

The well-known weighting schemes *Last-pooling* (with $M = 200$ frames), *Max-pooling* and *Mean-pooling* outperform *LSTM* showing that not only the last frame contains relevant information for the task. Among these three approaches, *Mean-pooling* achieves the best performance, and therefore, it seems better not to completely discard LSTM frames.

The *Logistic Regression Attention* method produces better results than the previous ones, although they are rather similar to *Mean-pooling* in the *RS* task. Nevertheless, it is clear that weighting the contribution of each frame can help to improve the performance of the system.

The first of our proposals, the use of weights derived from the Kalinli's normalized saliency map (*Normalized Saliency*) obtains better results than all previous approaches for the *RS* task and on average. For the *STP* and *SD* tasks, the achieved UARs are better than the previous strategies except for *Logistic Regression Attention* where results are similar. Nevertheless, our second approach, where the weights are extracted from the unnormalized saliency maps (*Unnormalized Saliency*), clearly

Table 2: *Unweighted Average Recalls (UARs) [%] for the baseline system and different LSTM-based classifiers for the Reading Sentence, Stroop Time Pressure and Stroop Dual tasks.*

System	Reading Sentence (RS)	Stroop Time Pressure (STP)	Stroop Dual (SD)	Average
SVM (baseline) [2]	61.50	66.70	56.90	61.60
LSTM	48.87 ± 1.36	55.42 ± 1.02	45.83 ± 4.09	49.61 ± 1.33
LSTM + VAD	45.34 ± 1.79	54.01 ± 2.02	46.60 ± 4.06	46.36 ± 1.51
LSTM Last-Pooling	52.42 ± 1.53	59.57 ± 2.81	46.60 ± 4.11	52.67 ± 1.30
LSTM Max-Pooling	59.87 ± 1.28	53.48 ± 0.98	41.95 ± 1.83	57.54 ± 1.18
LSTM Mean-Pooling	62.99 ± 0.82	60.69 ± 0.67	50.00 ± 2.07	61.61 ± 1.01
LSTM Logistic Regression Attention	63.58 ± 0.48	63.47 ± 0.67	54.59 ± 0.67	62.75 ± 0.59
LSTM Normalized Saliency	65.09 ± 0.55	63.06 ± 2.64	54.31 ± 2.01	63.86 ± 0.86
LSTM Unnormalized Saliency	66.97 ± 0.68	69.24 ± 0.52	63.69 ± 1.25	66.80 ± 0.50

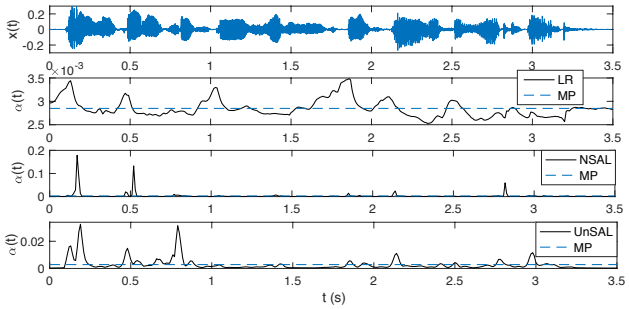

Figure 2: *Attention weights for one utterance belonging to the Reading Sentence task obtained with different strategies. MP: Mean-pooling; LR: Logistic regression attention; NSAL: Normalized saliency; UnSAL: Unnormalized saliency.*

outperforms all the rest of LSTM-based systems for all the tasks and on average. In particular, *Unnormalized Saliency* achieves a relative error reduction with respect to *Logistic Regression Attention* of 9.3%, 15.8%, 20.0% and 10.9% for the *RS*, *STP* and *SD* tasks and on average, respectively. These results corroborate our hypothesis that saliency signals could be used for establishing to some extent the relative importance of some frames for the CL level determination.

Fig. 2 depicts from top to bottom, the waveform of an utterance belonging to the *RS* task, the weights used in the *Logistic Regression Attention*, *Normalized Saliency* and *Unnormalized Saliency* approaches. Contrary to the observations made in [11], in our case, the regression attention weights are very uniform and closely resemble the mean-pooling weights (note the difference in scales). This justifies the fact that the results achieved by *Mean-pooling* and *Logistic Regression Attention* are rather similar. *Normalized Saliency* weights are very peaky, discarding many frames that could contain useful information for the task. However, weights of the *Unnormalized Saliency* approach presents a large degree of variation, suggesting that the unnormalized saliency signal becomes a good approximation of the amount of cognitive load content of a speech frame and it is useful in situations when not enough data is available for training more sophisticated attention models.

For comparison purposes, Table 3 contains the UARs achieved by several state-of-the-art systems on the CSLE database. As can be observed, our system obtains the second best result, only from behind [5], that was the winner of the 2014 COMPARE challenge for CL level classification.

Table 3: *Comparison results in terms of UAR [%] (average over the three tasks) for different approaches. SDC stands for Shifted Delta Coefficients and SCF for Spectral Centroid Frequency.*

System	Average
SVM (baseline) [2]	61.60
High-level features + SVM [30]	63.10
Fusion of [2], MFCC+SDC supervectors and SCF supervectors + GMM-SVM [3]	63.70
Feature Selection + Speaker Clustering + SVM [31]	64.80
LSTM Unnormalized Saliency (this paper)	66.80
Fusion of 4 Speech Streams + i-Vectors + SVM [5]	68.90

5. Conclusions and Future Work

In this paper, we have proposed an automatic system capable of classifying the cognitive load level of a speaker by analyzing his/her voice. We have presented two main contributions. Firstly, we have designed and evaluated for this task a LSTM-based system with different weighted pooling strategies. Secondly, we have proposed a novel attention mechanism based on the Kalinli's auditory saliency model. Experiments have shown that the weighted pooling LSTM system with weights derived from the unnormalized saliency maps achieves 13.5% and 10.9% relative error reductions with respect to the baseline SVM-based and the LSTM-based with logistic regression attention systems, respectively.

For future work, we plan to extend our research on CL level classification from speech in two directions: to analyze the relationship between the emphasized frames by the saliency model and some speech production properties associated with CL, such as speed rate, formant changes, etc. and, to study the use of alternative auditory saliency techniques [18].

6. Acknowledgements

The work leading to these results has been partly supported by Spanish Government grants TEC2017-84395-P and TEC2017-84593-C2-1-R. The authors would like to thank Prof. J. Epps for kindly providing the "Cognitive Load with Speech and EGG" (CSLE) dataset and Prof. B. Schuller and the rest of the 2014 COMPARE Challenge organizers for kindly providing the dataset partition and the baseline system.

7. References

[1] T. van Gog and F. Paas, *Encyclopedia of the Sciences of Learning*. Springer US, 2012, ch. Cognitive Load measurement, pp. 599–601.

[2] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load," in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association, September 14-18, Singapore, Proceedings*, 2014.

[3] J. M. K. Kua, V. Sethu, P. Le, and E. Ambikairajah, "The UNSW submission to INTERSPEECH 2014 compare cognitive load challenge," in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association, September 14-18, Singapore, Proceedings*, 2014, pp. 746–750.

[4] H. Boril, O. Sadjadi, T. Kleinschmidt, , and J. Hansen, "Analysis and detection of cognitive load and frustration in drivers speech," in *INTERSPEECH 2010 – 11th Annual Conference of the International Speech Communication Association, September 26-30, Makuhari, Chiba, Japan, Proceedings*, 2010, pp. 502–505.

[5] M. V. Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of cognitive load from speech using an i-vector framework," in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association, September 14-18, Singapore, Proceedings*, 2014, pp. 751–755.

[6] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks," in *ICASSP 2015 – 40th IEEE International Conference on Acoustics, Speech and Signal Processing, April 19-24, Brisbane, Australia, Proceedings*, 2015, pp. 4225–4229.

[7] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[8] R. Zazo, A. Lozano-Díez, J. González-Domínguez, D. T. Toledano, and J. González-Rodríguez, "Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks," *PLOS ONE*, vol. 11(1): e0146917, 2016.

[9] C. Huang and S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, September 8-12, San Francisco, USA, Proceedings*, 2016, pp. 1387–1391.

[10] ——, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *ICME 2017 – IEEE International Conference on Multimedia and Expo, July 10-14, Hong Kong, Proceedings*, 2017, pp. 583–588.

[11] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *ICASSP 2017 – 42th IEEE International Conference on Acoustics, Speech and Signal Processing, March 5-9, New Orleans, USA, Proceedings*, 2017, pp. 2227–2231.

[12] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *CVPR 2012 – 25th IEEE Conference on Computer Vision and Pattern Recognition, June 18-20, Rodhe Island, USA, Proceedings*, 2012, pp. 3642–3649.

[13] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *Journal of Machine Learning Research*, vol. 3, pp. 115–143, 2003.

[14] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *NIPS 2015 – 29th Conference on Neural Information Processing Systems, December 7-12, Montreal, Canada, Proceedings*, 2015, pp. 577–585.

[15] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv*, no. preprint arXiv:1508.04025, 2015.

[16] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv*, no. preprint arXiv:1706.00612, 2017.

[17] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 1009–1024, 2009.

[18] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Philosophical Transactions of the Royal Society B*, vol. 372, pp. 1–10, 2017.

[19] O. Kalinli, S. Sundaram, and S. Narayanan, "Saliency-driven unstructured acoustic scene classification using latent perceptual indexing," in *MMSP 2009 – IEEE 11th International Workshop on Multimedia Signal Processing, October 5-9, Rio de Janeiro, Brazil, Proceedings*, 2009, pp. 1–6.

[20] T. F. Yap, *Speech production under cognitive load: Effects and classification*. Ph.D. dissertation, The University of New South Wales, Sydney, Australia, 2012.

[21] J. R. Stroop, "Studies of interference in serial verbal reactions," *Journal of Experimental Psychology*, vol. 18, no. 6, 1935.

[22] F. Eyben, F. Weninger, F. Groβ, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *MM 2013 – 21st ACM International Conference on Multimedia, October 21-25, Barcelona, Spain, Proceedings*, 2013, pp. 835–838.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, no. 11, 2009.

[24] M. Abadi *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org, 2015.

[25] F. Chollet *et al.*, *Keras: the Python Deep Learning library*. Software available from https://github.com/fchollet/keras, 2015.

[26] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python," in *SciPy 2015 – 14th Python in Science Conference, July 6-12, Texas, USA, Proceedings*, 2015, pp. 18–25.

[27] E. Macaluso, *MT_TOOLS: Computation of Saliency and Feature-Specific Maps*. Software available from http://www.brainreality.eu/mt_tools/, 2010.

[28] A. Berthold and A. Jameson, "Interpreting symptoms of cognitive load in speech input," in *UM 1999 – 7th International Conference on User Modeling, June 20-24, Banff, Canada, Proceedings*, 1999, pp. 235–244.

[29] C. Muller, B. Grossmann-Hutter, A. Jameson, R. Jummer, and F. Wittig, *Lecture Notes in Computer Science*. Springer, 2001, ch. Recognizing time pressure and cognitive load on the basis of speech: an experimental study, pp. 24–33.

[30] C. Montacié and M. J. Caraty, "High-level speech event analysis for cognitive load classification," in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association, September 14-18, Singapore, Proceedings*, 2014, pp. 731–735.

[31] G. Gosztolya and L. Tóth, "A feature selection-based speaker clustering method for paralinguistic tasks," *Pattern Analysis and Applications*, vol. 21, no. 1, pp. 193–204, 2018.

[32] *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association, September 14-18, Singapore, Proceedings*, 2014.