Affective Burst Detection from Speech using Kernel-fusion Dilated Convolutional Neural Networks

Berkay Köprü and Engin Erzin
KUIS-AI Laboratory, Multimedia, Vision and Graphics Group
College of Engineering, Koç University, Istanbul, Turkey
bkopru17,eerzin@ku.edu.tr

***Abstract*—As speech interfaces are getting richer and widespread, speech emotion recognition promises more attractive applications. In the continuous emotion recognition (CER) problem, tracking changes across affective states is an essential and desired capability. Although CER studies widely use correlation metrics in evaluations, these metrics do not always capture all the high-intensity changes in the affective domain. In this paper, we define a novel affective burst detection problem to capture high-intensity changes of the affective attributes accurately. We formulate a two-class classification approach to isolate affective burst regions over the affective state contour for this problem. The proposed classifier is a kernel-fusion dilated convolutional neural network (KFDCNN) architecture driven by speech spectral features to segment the affective attribute contour into idle and burst sections. Experimental evaluations are performed on the RECOLA and CreativeIT datasets. The proposed KFDCNN outperforms baseline feedforward neural networks on both datasets.**

***Index Terms*—Emotion recognition, affective burst detection, kernel fusion, convolutional neural networks, speech analysis**

I. Introduction

Emotions in humans are driven by biological stimuli from external or internal stimuli and interact with the cognition system in the brain [1]. Since the emotional system drives human behaviors, automatic recognition of emotions has become an attractive and significant research field in the last two decades.

Speech Emotion Recognition (SER) has drawn significant attention in the recent literature as speech is an easy-to-collect modality and presents a rich and robust representation of emotion under clean acoustic conditions [2], [3], [4], [5]. SER also keeps attracting attention with the surge of speech interfaces in the Internet-of-Things applications [2].

Discrete emotion recognition (DER) is a subclass of SER, which focuses on the categorical representation of emotions such as anger and happiness. In DER studies, an audio signal is widely represented with low-level descriptors (LLDs), such as pitch, energy, zero-crossing rate, and spectral features [6], [7]. The classification task in DER has been addressed using LLD features by hidden Markov model-based approaches as in [6] and also by recurrent neural networks (RNNs) based approaches as in [7]. Alternatively in [8], raw speech is processed with multiple 1-dimensional convolutions to classify emotions.

Alternatively, continuous emotion recognition (CER) represents emotions in a 3-dimensional continuous affect attribute space whose dimensions are Arousal, Valence and Dominance, respectively representing activeness - passiveness, positiveness - negativeness and dominance-submissiveness [9]. CER studies use both tailored features as in [10], [11] or learned features as in [12], [13], [14]. In [10], 23 LLDs from eGeMAPS [15] are extracted from the audio signal. Then, based on these features, a stacked long short-term memory (LSTM)-RNNs model is proposed for CER. In [11], the audio signal is represented by a combination of the Mel-frequency Cepstral Coefficients (MFCCs), delta and acceleration of MFCCs. Then using multi-task learning, Arousal, Valence, and Dominance attributes are estimated in parallel. Trigeorgis et al. propose a convolutional recurrent neural network where two convolutional layers acted as a learned feature extractor on the raw audio signal [12]. Similarly, Tzirakis et al. adopt convolutional neural networks (CNNs) to produce audio embeddings for CER [13]. In CER studies, correlation-based metrics are widely used for evaluation tasks since the trend of the predicted attribute is more important than the actual level of the prediction. On the other hand, correlation metrics do not always grant effective capturing of all high-intensity changes in the affective domain.

Due to the categorized nature of the DER, changes between the categories (inter-emotion transitions) can be observed, however transitions within each category (intra-emotion transitions) do not appear or are not typically available for the DER studies. For instance, while On the other hand, continuous affect attributes in the CER problem provide the necessary intensity fluctuations for the inter-emotion and intra-emotion transitions.

The detection of inter- and intra-emotion transitions, i.e., affective change detection, has a significant importance, and it is widely studied in the psychology domain under the *mismatch negativity* (MMN) literature [16], [17], [18].

Although the information processing capacity of human beings is limited, people do attend emotional changes [19]. Hence it is essential and valuable to detect changes in the emotional domain for fair HCI applications. As the emotional expressions drive the communication for possible threats in the environment, even when attention is engaged in a concurrent task, emotional information is prioritized and automatically

processed [20]. Hence, following changes in the affective domain can be critically important to design natural human-computer interaction applications.

Affective change detection is studied in the DER context by [21] and CER context by [22]. *Affective change points* are defined as the transition points between the emotions in [21]. These points are estimated using a Gaussian mixture model based architecture with and without prior emotion class information. In [22], emotional hotspots are defined as sections deviated from the median of the affective attribute. They proposed a qualitative agreement-based assessment method to map affective attributes into low, high, neutral, and non-consensus sections. A section is labeled as a low if that section is under the median, and high if it is above the median. Then, bidirectional long short-term memory (BLSTM) is operated on 88 eGeMAPS [15] features, which are extracted from acoustic signals, to classify the trend. However, this approach highlights flat sections that are deviated from the median, and it resembles a quantization approach more than tracking the high-intensity regions as the labeling procedure misses all the inter- and some of the intra-emotion transition regions in the affective domain.

In this study, we address the segmentation of the affect contour into affective burst and idle regions. Unlike [22], we label sections regarding the intensity of change. Affective burst sections were then estimated using spectral features of speech as input and a kernel-fusion dilated convolutional neural network (KFDCNN) as the classier. To summarize, the main contributions of this study are as follows:

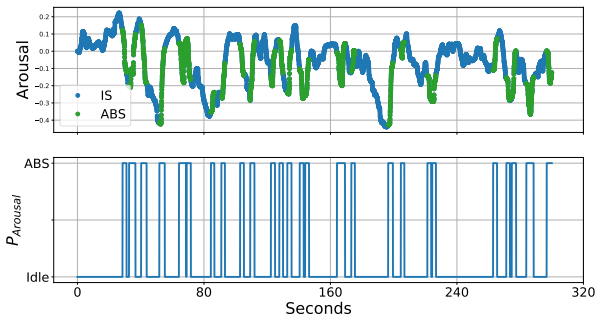
- We propose a new labeling mechanism to define the affect contour's high intensity and idle segments.
- We formulate a novel affective burst detection problem capturing the high-intensity changes, which can improve the understanding of the inter-emotion and intra-emotion transitions in the scene.
- We propose a novel architecture KFDCNN for affective burst detection from speech.
- We evaluate the RECOLA and CreativeIT datasets with classification metrics.

II. Method

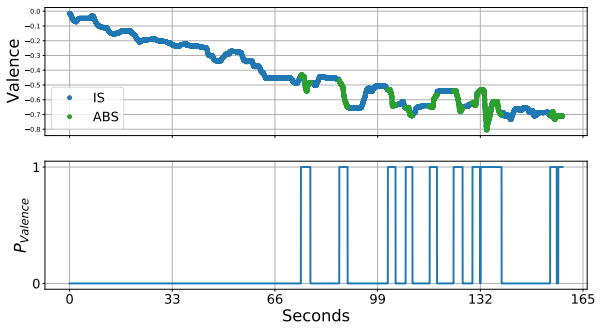
We propose a convolutional neural network (CNN) architecture where parallel convolutions process the input with different dilated kernel lengths to detect affective bursts. In this section, the affective burst detection problem is first defined, then feature extraction from speech signals is described. Later, an affective burst detection framework based on kernel fusion and CNN's is described.

A. Affective Bursts

Affective burst detection is a two-class classification problem where the affective contour is segmented into affective burst and idle classes. We define the affective burst segment (ABS) as the region in which the affect attribute contour is changing rapidly with high gradients. Respectively, idle segments (ISs) cover the complement of the ABSs and correspond



(a) Labeled affective burst segments (ABS) for Arousal from RECOLA dataset



(b) Labeled affective burst segments (ABS) for Valence from CreativeIT dataset

Fig. 1. Sample Arousal and Valence contours with affective burst and idle segment labelings calculated using Equation 3

to the regions where the affect attribute contour is changing slowly with low gradients.

The ground-truth ABS annotations are generated in two steps. First affective burst points (ABPs) on the affect attribute contours are detected, then these points are then extended into segments referred to as ABSs. Note that all non-ABS regions are referred to as idle segments.

Affective burst points (ABPs) are set based on the first-order regression coefficients of the arousal and valence attributes as

$$d_e[n] = \frac{\sum_{l=1}^{L}(e[n+l] - e[n-l])l}{2\sum_{l=1}^{L} l^2} \tag{1}$$

where $d_e[n]$ is the delta coefficient of attribute e (Arousal or Valence) at sample index n. By selecting a threshold τ_e, we can define the ABP indicator function p_e at sample index n as

$$p_e[n] = \begin{cases} 0 & \text{if } \tau_e < |d_e[n]| \\ 1 & \text{if } |d_e[n]| \geq \tau_e. \end{cases} \tag{2}$$

Then the ground truth binary segment labels are extracted as

$$P_e[n+i] = \begin{cases} 1 & \forall n \text{ s.t. } p_e[n] = 1 \text{ and } i = -\Delta, \ldots, \Delta \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where an ABS of temporal size $w_s = 2\Delta + 1$, where Δ is the defining parameter of temporal extend in forward and

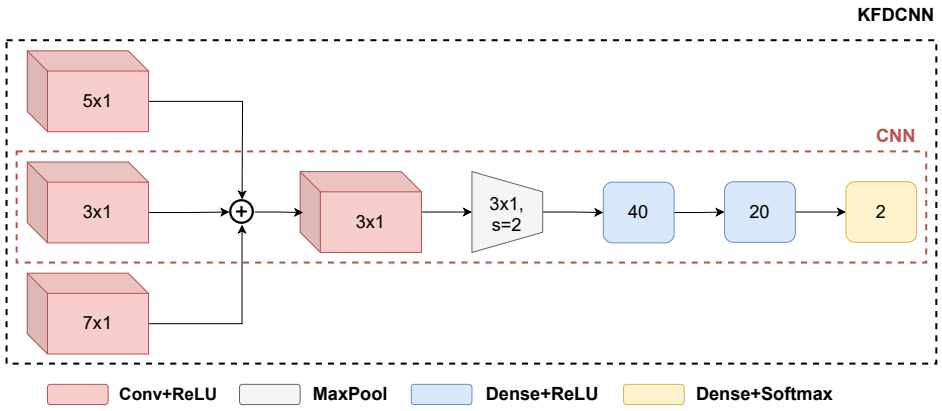


Fig. 2. Kernel fusion dilated convolutional neural network for affective burst detection, where the convolutional layers at the input have a dilation rate of 5 **(see Section II-C for explanation)**

backward direction, is centered for each ABP on the indicator function $p_e[n]$. Note that the resulting size of ABSs can be longer than w_s when two consecutive ABPs are closer than w_s. Sample affect attribute contours for arousal and valence, indicated as ABS and IS in different colors, together with $P_e[n]$ are depicted in Figure 1.

B. Feature Extraction

In this study, we use the *extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)* representing the spectral and temporal characterization of speech signal for ABS detection [15]. The 88-dimensional eGeMAPS features are calculated as statistics of 25 *low-level descriptors* (LLDs), such as Mel-frequency Cepstral Coefficients, Pitch, and Loudness, using the OpenSMILE toolkit [23].

The LLDs are extracted using a window size of 20 ms, and the 88-dimensional eGeMAPS features are calculated over 500 ms with a hop duration of 40 ms. Hence the eGeMAPS features are extracted at 25 fps and represented at time frame n as $F[n] \in \mathbf{R}^{88}$.

C. Kernel Fusion Convolutional Neural Network

The proposed KFDCNN architecture is depicted in Figure 2. KFDCNN comprises both dilated parallel and typical convolutional layers, max-pooling layer, and multiple fully connected layers. The kernel fusion layer at KFDCNN includes dilated convolutional kernels with different lengths that help learn temporal relations in different resolutions. Hence, this layer enriches representation for the ABS detection.

Furthermore, with dilated kernels, each layer has longer receptive fields, resulting in convolution outputs that capture long-term information, which is especially important for slowly varying emotional processes.

The proposed KFDCNN architecture processes a window of features which is represented as $I[n] = \{F[n - T], F[\, n - T + s], \ldots, F[n+T-s], F[n+T]\}$ where s is the dilation rate, $2T$ represents the receptive field size at the input, and $I[n] \in \mathbf{R}^{\frac{2T}{s} \times 88}$. Then KFDCNN outputs $\hat{y}[n] \in \mathbf{R}^2$ from the input $I[n]$.

D. Model Training

The architecture is trained for ABS detection through a binary classification task. As the training scheme is the same for both attributes, for the sake of simplicity, we will drop the attribute indicator e from the P and other variables that are introduced in Equation 4. However, this task has an imbalanced nature. There are a small number of sections labeled as ABS, while a high number of sections are labeled as IS. To overcome the imbalance problem, we adopted weighted negative log-likelihood ratio loss:

$$L(\mathbf{P}, \hat{\mathbf{y}}, \theta) = \frac{1}{(\theta_0 + \theta_1)} \sum_n (-\theta_{P[n]} \log(\hat{y}_{P[n]}[n])) \quad (4)$$

where θ_i is the weight for class i, $P[n] \in \{0, 1\}$ is the binary segment label at time frame n, $\hat{y}_{P[n]}[n]$ is the probability output of KFDCNN corresponding to the segment label $P[n]$ at frame n, where $\hat{y}[n] = \{\hat{y}_0[n], \hat{y}_1[n]\}$. The class weight θ_i is extracted as

$$\theta_i = \frac{1}{\text{frequency of } i^{th} \text{ class}} \quad (5)$$

for class index $i = 0, 1$ corresponding to the IS and ABS.

III. Experimental Evaluations

The proposed architecture in Section 2 is evaluated on the RECOLA [24] and the CreativeIT [25] datasets. In this section, datasets and implementation details are introduced. Then evaluation metrics are described. Finally, the performance of the KFDCNN is compared against a baseline feed-forward neural network (FFN) together with the CNN and the dilated CNN (DCNN).

A. Datasets

We train and evaluate the ABS detection task on the widely used multi-modal datasets RECOLA and CreativeIT. The RECOLA dataset is composed of multi-modal recordings of dyadic conversations of 27 French speakers. As a part of the AVEC16 challenge, the dataset is divided into uniform-sized training, development, and test sets. While annotations

for the training and the development sets are available, the annotations for the test set are not public. Publicly available annotations are for the arousal and valence attributes at 25 Hz rate.

USC CreativeIT is a multi-modal database of theatrical improvisations. Each interaction, on average, has a length of 3.5 minutes and is captured by recordings of the body Euler angles and speech from the participants. The dataset includes references for arousal, valence, and dominance. It is divided into five sessions which are mutually exclusive in terms of speakers. The cross-validation procedure on this dataset is held by leave-one session out to preserve speaker independence.

TABLE I
STATISTICS OVER THE RECOLA AND CREATIVEIT DATASETS AFTER THE GROUND-TRUTH ABS ANNOTATIONS ON AROUSAL (A) AND VALENCE (V) CONTOURS: NUMBER OF ABSS, MEAN DURATION OF ABSS, TOTAL DURATION OF ABSS, AND MEAN ABSOLUTE DELTA ($|d|$) OF THE ABSS

Stats	**RECOLA**		**CreativeIT**	
	A	V	A	V
# ABSs	446	461	641	632
Mean ABS dur (sec)	3.4	3.8	3.6	3.6
Total ABS dur (sec)	1510	1744	2308	2296
Mean $\|d\|$ of ABSs	0.0030	0.0017	0.0014	0.0009
Total dur (sec)	5400	5400	7708	7708

Table I presents the statistical characterization of the ground-truth ABS annotations on the arousal and valence contours of the RECOLA and CreativeIT datasets. Total ABS region durations cover around 30% of the datasets with similar mean ABS durations of 3.6 seconds. Mean absolute delta ($|d|$) values for the ABS regions are observed higher for the RECOLA dataset, indicating higher affect contour changes for the RECOLA.

B. Implementation Details and Setup

Experimental evaluations of the proposed ABS detection system are executed using cross-validation. For the RECOLA dataset, two videos are selected as the test set, and the rest are chosen for training and validation, resulting in 9 folds. On the other hand, we apply a leave-one-session-out strategy for the CreativeIT dataset, for each fold, a session is chosen as a test set, and the rest are used for training and validation, resulting in 5 folds.

We set the length, L, of the first-order regression coefficients to capture 0.8 seconds temporally, the ABS temporal window size, w_s, is set to span 2 seconds. In (2), we set two thresholds, τ, values, one for each dataset, so to cover 30% of the datasets as the ABS regions.

The input of the KFDCNN, $I[n]$, is set with $T = 100$ and $s = 5$ which spans an 8 seconds temporal window with dilation rate 5. Kernel fusion layer at KFDCNN has three parallel 1-dimensional convolutions with kernels sizes 3, 5, and 7, and these kernels have a dilation rate of $s = 5$. The second 1-dimensional convolution has a kernel length of 3 with a dilation rate of 1. The max-pool layer down-samples the temporal dimension in half. The output of the max-pool layer is flattened from 2-dimension into 1-dimension and fed into fully connected layers with node sizes of 40, 20, and 2, respectively.

Single kernel and no dilation derivatives of the KFDCNN are also defined and evaluated to assess the proposed model's performance better. The CNN architecture, which is depicted within the red dashed lines in Figure 2, has a single kernel set with a size of 3 and a dilation rate of 1. In order to have comparable complexity with the KFDCNN, the input feature of the CNN is set with $T = 20$ and $s = 1$, which spans a 1.6 seconds long temporal window without dilation. A dilated CNN (DCNN) architecture is also defined by setting the input feature representation with $T = 100$ and $s = 5$.

C. Affective Change Point Detection Performances

Unweighted average F-score (UAF1) and Recall (UAR) metrics are computed at frame level via cross-validation and used for the performance evaluations.

Table II presents F1-score and Recall performances of the KFDCNN against the baselines (SVM and FFN), CNN, and DCNN. CNN-based architectures outperform the baseline in all comparison metrics by at least 2 percent-point (pp) at RECOLA and CreativeIT databases. This result stresses the importance of temporal information for the detection of ABSs. KFDCNN distinctly performs better than CNN and DCNN models among the CNN-based architectures. Performance improvement for the KFDCNN is most significant for arousal and valence in the RECOLA dataset, while only for valence in the CreativeIT dataset.

Comparing CNN with DCNN, dilation improves the F-score performance by 5 pp for arousal and 3 pp for valence in the RECOLA dataset. Moreover, similar improvements are also seen with the CreativeIT dataset. Significant temporal context due to dilated kernels is crucial to differentiate idle sections from ABS. This observation supports that consecutive temporal features carry less extrinsic information due to the slowly varying nature of emotional processes.

Comparing DCNN with KFDCNN, kernel fusion improves F-score approximately by 3 pp for arousal and 2 pp for valence at the RECOLA database. Similarly, at the CreativeIT database, improvements are 2 pp for valence. On the other hand, DCNN has only 0.2 pp better performance for arousal at the CreativeIT database. This consistent improvement indicates that learning relationships at different temporal resolutions improve ABS detection.

KFDCNN has superior performance on the RECOLA than CreativeIT, by approximately 10 pp for arousal and 3 pp for valence. The change in the performance could be because RECOLA is not an acted dataset, and as a result, it includes more spontaneous changes and exhibits higher mean absolute delta, $|d|$, values within ABSs compared to the CreativeIT.

IV. CONCLUSION

This study presents affective burst detection as a crucial affective computing problem that can capture affective fluctuations better in the inter-and intra-emotion domain. We address

TABLE II
F-SCORE AND RECALL PERFORMANCE RESULTS OF THE BASELINE (FFN), CONVOLUTIONAL NEURAL NETWORK (CNN), DILATED CONVOLUTIONAL NEURAL NETWORK (DCNN), AND THE PROPOSED KFDCNN FRAMEWORK FOR THE AFFECTIVE BURST DETECTION OVER THE RECOLA AND CREATIVEIT DATASETS ON AROUSAL (A) AND VALENCE (V) CONTOURS

Model	**RECOLA**				**CreativeIT**			
	UAF1		UAR		UAF1		UAR	
	A	V	A	V	A	V	A	V
SVM	51.3	52.3	52.0	54.6	45.0	50.9	53.6	57.7
FFN	56.2	53.7	56.8	54.4	48.1	51.0	55.8	56.1
CNN	59.0	56.3	60.1	59.0	53.8	56.4	58.5	59.6
DCNN	64.3	59.4	64.8	60.3	**57.2**	56.8	**60.3**	58.0
KFDCNN	**67.0**	**61.4**	**68.5**	**62.2**	57.0	**58.5**	59.4	**60.0**

the affective burst detection as an imbalanced binary problem of segmenting affective contour into burst and idle regions.

First, we label the affective contour by first detecting ABPs over the derivatives of the affective attributes. Later, the annotations are generated by extending the ABPs into segment vector ABSs. The proposed KFDCNN architecture is trained with the generated annotation targets. The conducted experiments show that the KFDCNN outperforms the baseline architecture for F1-score and Recall on both RECOLA and CreativeIT datasets. Moreover, we depicted the importance of dilation and kernel fusion by comparing KFDCNN with CNN and DCNN. It is seen that larger receptive field size due to dilation brings at least 3 pp improvements, and kernel fusion brings at least 2 pp improvements. Considering these observations, we suggest using KFDCNN for the affective burst detection.

REFERENCES

[1] Joseph E. Ledoux, "Cognitive-emotional interactions in the brain," *Cognition and Emotion*, vol. 3, no. 4, pp. 267–289, 1989.

[2] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and I Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artif Intell Rev*, vol. 43, pp. 155–177, 2015.

[3] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[4] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, and Yuan Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, oct 2018.

[5] Ligang Zhang, Brijesh Verma, Dian Tjondronegoro, and Vinod Chandran, "Facial expression analysis under partial occlusion: A survey," *ACM Computing Surveys*, vol. 51, no. 2, pp. 1–49, apr 2018.

[6] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Hidden Markov model-based speech emotion recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2003, vol. 2, pp. 1–4.

[7] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, jun 2017, pp. 2227–2231.

[8] Puneet Kumar, Sidharth Jain, Balasubramanian Raman, Partha Pratim Roy, and Masakazu Iwamura, "End-to-end Triplet Loss based Emotion Embedding System for Speech Emotion Recognition," in *The 25th International Conference on Pattern Recognition (ICPR 2020)*, 2020.

[9] Harold Schlosberg, "Three dimensions of emotion," *Psychological Review*, vol. 61, no. 2, pp. 81–88, mar 1954.

[10] Maximilian Schmitt, Nicholas Cummins, and Björn W. Schuller, "Continuous Emotion Recognition in Speech — Do We Need Recurrence?," in *Interspeech 2019*, ISCA, sep 2019, pp. 2808–2812, ISCA.

[11] Berkay Köprü and Engin Erzin, "Multimodal continuous emotion recognition using deep multi-task learning with correlation loss," *arXiv preprint arXiv:2011.00876*, 2020.

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

[13] Panagiotis Tzirakis, Jiaxin Chen, Stefanos Zafeiriou, and Björn Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, apr 2021.

[14] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, sep 2018, vol. 2018-April, pp. 5089–5093.

[15] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[16] Alie G. Male, Robert P. O'Shea, Erich Schröger, Dagmar Müller, Urte Roeber, and Andreas Widmann, "The quest for the genuine visual mismatch negativity (vMMN): Event-related potential indications of deviance detection for low-level visual features," *Psychophysiology*, vol. 57, no. 6, pp. e13576, jun 2020.

[17] Klara Kovarski, Marianne Latinus, Judith Charpentier, Helen Cléry, Sylvie Roux, Emmanuelle Houy-Durand, Agathe Saby, Frédérique Bonnet-Brilhault, Magali Batty, and Marie Gomot, "Facial Expression Related vMMN: Disentangling Emotional from Neutral Change Detection," *Frontiers in Human Neuroscience*, vol. 11, pp. 18, jan 2017.

[18] Petia Kojouharova, Domonkos File, István Sulykos, and István Czigler, "Visual mismatch negativity and stimulus-specific adaptation: the role of stimulus complexity," *Experimental Brain Research*, vol. 237, no. 5, pp. 1179–1194, may 2019.

[19] Maurizio Corbetta and Gordon L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Reviews Neuroscience*, vol. 3, no. 3, pp. 201–215, 2002.

[20] J.A. Hinojosa, F. Mercado, and L. Carretié, "N170 sensitivity to facial expression: A meta-analysis," *Neuroscience & Biobehavioral Reviews*, vol. 55, pp. 498–509, 2015.

[21] Zhaocheng Huang, Julien Epps, and Eliathamby Ambikairajah, "An Investigation of Emotion Change Detection from Speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] Srinivas Parthasarathy and Carlos Busso, "Predicting Emotionally Salient Regions using Qualitative Agreement of Deep Neural Network Regressors," *IEEE Transactions on Affective Computing*, 2018.

[23] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, New York, NY, USA, 2010, MM '10, p. 1459–1462, Association for Computing Machinery.

[24] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, 2013.

[25] Angeliki Metallinou, Zhaojun Yang, Chi-chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language Resources and Evaluation*, vol. 50, no. 3, pp. 497–521, 2016.