SMM19, Workshop on Speech, Music and Mind 2019
14 September 2019, Vienna, Austria

Detection of emotional states of OCD patients in an exposure-response prevention therapy scenario

Kaajal Gupta[1], Anzar Zulfiqar[2], Pushpa Ramu[2], Tilak Purohit[3], V. Ramasubramanian[3]

[1]The International School Bangalore (TISB), Bangalore, India
[2]Samsung R&D Institute, Bangalore (SRIB), India
[3]International Institute of Information Technology - Bangalore (IIIT-B), Bangalore, India

gkaajal@tisb.ac.in, {anzar.zulfi, pushpa.r, tilak.purohit,}@iiitb.org,
v.ramasubramanian@iiitb.ac.in

Abstract

We address the problem of detection of emotional states of obsessive-compulsive disorder (OCD) patients in an exposure-response prevention (ERP) therapy protocol scenario. Here, it is required to identify the emotional levels of a patient at a granular level needed for successful progression of the therapy, and one of the major hurdles in this is the so called alexithymia (subclinical inability to identify emotions in the self). Alternately, we propose estimating the emotional state of an OCD patient automatically from raw speech signal, elicited under a situation-based emotion entry to an on-line therapy aid. Towards this, we propose a novel multi-temporal CNN architecture for end-to-end 'speech emotion recognition' (SER) from raw speech signal. The proposed architecture allows for multiple time-frequency resolutions with multiple filter banks having different time-frequency resolutions to create feature-maps (ranging from very narrow-band to very wide-band spectrographic maps in steps of fine time-frequency resolutions). On SER task, we show 2-8% absolute enhancement in accuracy for the multi-temporal cases (e.g. 3, 6 branches) over the conventional single-temporal CNNs. As a position paper, we identify further work as fine-granular emotion detection of the OCD emotional states via a valence-arousal-dominance detection to derive the 'degree' of emotion of an OCD patient.

Index Terms: OCD mental states, emotional states, multi-temporal CNN, end-to-end speech emotion recognition

1. Introduction

Speech emotion recognition (SER) [1] has attracted considerable attention for nearly 2 decades with several promising results and state of art performances. SER is typically called for in various application domains such as audio-based multimedia (e.g. movie) content indexing, call center analytics (to determine the emotional state of a caller), rich transcription of various speech data, spoken dialog systems to detect and track the emotional state of an user etc. In this paper, we address the problem of detecting and tracking the emotional state of OCD patients from raw speech signal in an exposure-response prevention (ERP) therapy protocol scenario. This scenario, conventionally uses a qualitative assessement of the patients anxiety level, but suffers from the difficulty the patients face in being able to quantify their anxiety. This is especially challenging for OCD sufferers. In this work, we aim to quantify this assessment and measurement of anxiety and emotional state of the OCD patient through an on-line protocol, which makes available raw speech elicited from the patient and allows a SER system to detect and track the emotional state of the patient at a high granularity expected to be possible from a 3-D model of valence-arousal-dominance scale.

In this paper, we propose a novel multi-temporal CNN architecture for end-to-end 'speech emotion recognition' (SER) from raw speech signal focusing on a specific aspect of the CNNs, namely, the kernel sizes used in the convolutional kernels, and point out that for applying CNNs on raw 1-dimensional signals such as speech-, audio- and music-waveforms, it becomes important to 'provide' for a variable kernel size, to exploit and resolve the well known time-frequency trade-off inherent in such 1-dimensional convolution (or windowed linear filtering) operation. While this applies to 2-dimensional images also, this issue of having to address the time-frequency trade-off in the application of a filter-bank kind of operation (what a set of kernels in a CNN layer do) has been more or less overlooked in the image-CNN community, and even more so in 1-d signal processing, where it applies more readily. Here, we apply this architecture for the SER problem and our focus and contributions are along the following lines:

1. To show the very significant performance gain (2-8% absolute) by the multi-temporal architecture (with 6 branches) over a conventional single-branch CNN.
2. As a position paper, we propose to adapt this architecture for detecting and tracking emotional states of OCD patients (in an on-line therapy protocol), leveraging its enhanced performance potential for valence-arousal detection to map to an emotional category and 'degree' of emotion in a fine-grained emotional state detection.

2. Situation-based emotion entry for OCD patients

2.1. Obsessive-Compulsive Disorder

Obsessive-compulsive disorder is a common and highly impairing mental disorder, considered to be one of the most debilitating psychiatric illnesses. It is characterized by distressing thoughts and repetitive behaviors that are interfering, time-consuming, and difficult to control [2].

Treatment for obsessive-compulsive disorder is comprised of Exposure-Response Prevention (ERP) therapy which is a type of Cognitive-Behavioral Therapy (CBT). Cognitive therapy guides a patient in identifying and modifying patterns of thoughts and behaviors that cause anxiety and distress. ERP involves the patient deliberately exposing themselves to the triggers of their obsessive thoughts. The goal is to normalize the triggers for the patient, and in turn modify their response to them, reducing the frequency of compulsions and severity of obsessions. [3]

10.21437/SMM.2019-5

Significant reduction in OCD symptoms was observed for 80% of patients undergoing ERP [4]. The therapy is conducted by the therapist on an outpatient basis once a week with 'homework for the patient, which may consist of daily exposures to be completed in between therapy sessions. Compliance with such homework sessions is strongly correlated with recovery from OCD, as can be seen in numerous studies where the fall in YBOCS (Yale-Brown Obsessive-Compulsive Scale) score is closely interlinked with homework compliance [5], [6].

2.2. Need for an online self-help app

To ensure that homework is being completed and reported to the therapist accurately, an online app that provides a collection of necessary exercises and sends the information to the therapist would be useful. Liberate: My OCD Fighter was developed for this purpose (Fig.1). The app also helps the patient learn more about OCD, track their progress, and provides information the methods to combat OCD. In addition, it contains exercises with tips for ERP and CBT, which allow the user and therapist to track the progress made. This is expected to improve patient compliance, as the therapist can confirm that the user is, in fact, doing their homework exercises.

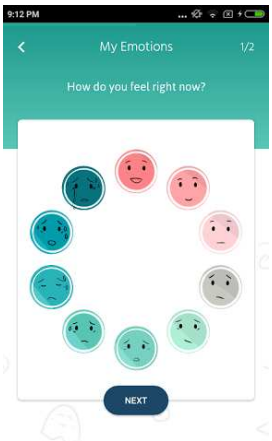

Figure 1: *A typical interaction state in the 'My OCD Fighter' app allowing the OCD patient to qualitatively enter his emotional state*

2.3. Tracking the emotional state of users

2.3.1. Motivation

Effectiveness of ERP is typically measured based on patient anxiety level recorded at periodic intervals in clinics [7], which includes qualitative emotions of the patient. Measurement is based on, direct interaction between the therapist and patient, with the patient rating their anxiety on a scale of 1 to 10. This tracking allows the therapist to determine the progress made by the patient with ERP and future course of action.

An obstacle faced by users with this premise is the difficulty in being able to quantify their anxiety. This is especially challenging for OCD sufferers.

Research has shown strong positive correlation between alexithymia (subclinical inability to identify and describe emotions in the self) and OCD [8], [9] making it harder for users to identify their emotional levels at a granular level, which is needed for successful progression of their therapy. This is the motivation behind our research into providing an alternate method to estimate the anxiety and emotional state of an OCD patient during ERP.

2.3.2. Exposure-Response Prevention and Situational Emotion Entries

Exposure therapy is typically practiced through a fear ladder. A fear ladder is composed of a list of the triggers that cause anxiety-provoking obsessive thoughts and thus compulsive urges in the patient [10]. The triggers are ordered by the level of anxiety they cause the patient. This allows the patient to progressively expose themselves to their triggers in ascending order of the anxiety caused by it.

Progression over the fear ladder is determined based on changes in patients anxiety levels after exposure. For instance, a person with contamination OCD may be shown an image of dirty tap. The patient may be asked to voice their feelings while viewing the image. The level of anxiety is measured based on their response and emotions identified from their voice. This is a less intense form of exposure for the user, as they are not prevented from completing their compulsion. Moreover, since this exercise is performed daily unlike ERP exercises, the patients therapist can chronologically view how the patients OCD has improved or worsened from their response to these triggers.

2.3.3. Mechanism

The fear ladder (Fig. 2) is composed of ten steps of exercises with increasing levels of difficulty for the patient in terms of the anxiety/distress induced by them. The patient is expected to start at step 1. Based on the type of OCD that they suffer from they select a trigger, set the amount of time for the exercise and begin the exposure. The app records the anxiety and emotional state of the patient every ten minutes. Successful completion of a step is defined as decrease in the user anxiety by at least 5 degrees between the start and end of the exercise. The user anxiety is defined on the scale of 1 to 10. The user accesses next steps upon successful completion of the previous steps.

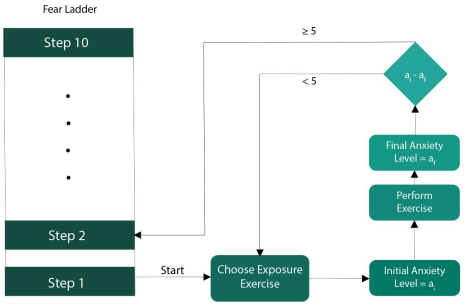

Figure 2: *The fear ladder and interactive therapy steps to traverse the ladder*

2.4. Emotional states

Accurate estimation of the emotions of the patient is essential for the success of Exposure Therapy. The current model determines emotions from raw speech, and the emotions are classified into primary or baseline emotions. Similar to the (2-d) circumplex model of affect by Russel [11], Plutchiks wheel of emotions (a 3-d circumplex model) [12], [13] maps the primary emotions (via certain combinations) to secondary emotions.

Plutchik considered 8 primary emotions: happiness, sadness, fear, disgust, anger, surprise, anticipation and trust. The secondary and tertiary dyads are considered to be combinations of these baseline emotions. For instance, the combination of anticipation and trust forms hope and the combination of anger and disgust forms contempt. There is a total of 56 emotion combinations possible at a single intensity level [14]. This model

can be used to extract a wider array of emotions from the emotions derived from the raw speech. The secondary emotions such as guilt (a combination of joy and fear), or shame (a combination of fear and disgust) would be extremely useful from the ERP progression point of view.

Further, the cones vertical dimension represents varying intensities of emotions; for instance, joy begins with serenity, and intensifies into ecstasy [13]. The emotion intensity estimation derived from speech can be fine-tuned by information such as; the values of valence (in relation to the concept of polarity), arousal (a calm-excited scale) and dominance (perceived degree of control in a (social) situation) [15]. The average value of the valence, arousal and dominance of discrete emotions in the 3D emotion space has been determined, and can be compared with respective values of the raw speech [16]. For example, the valence, arousal and dominance for 'anger' was measured to be -0.35 0.17, 0.46 0.18 and 0.53 0.14. Anger is found to be very negative (low valence), very excited (high arousal) and very strong (high dominance). The emotional space spanned by the valence-arousal-dominance model is shown in Fig. 3 [15].

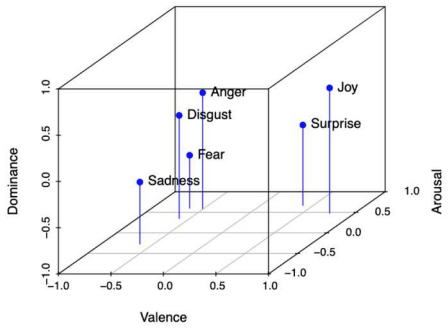

Figure 3: *Emotional space spanned by the valence-arousal-dominance model*

Returning to the case of anger, if the valence of speech is lower than the average, arousal is greater than the average, and dominance is greater than the average, it can be inferred that anger is of a higher degree and can be classified as rage from Plutchiks model. This logic can be replicated for different primary emotions to derive the secondary emotions.

2.4.1. Usage of Collected Information

The information on the change in emotional state and anxiety of the patient is sent to the therapist on a weekly basis in the form of progress reports. These progress reports will, as a result, become more comprehensive as the extraction of emotions from a patients voice can identify emotions the patients had not themselves recognized. The therapist and patient can together analyze the cause of each emotion, and devise new exposures if there is a negligible change in the degree of anxiety of the patient before and after exposure.

3. Multi-temporal CNN architecture

The multi-temporal CNN architecture considered here is as shown in Fig. 4, comprising two parts (as in Fig. 4a) and b)): a) Formation of the multi time-frequency spectrographic feature maps and b) From the feature maps to fully connected layers. These are described in details below.

Fig. 4a) is the essential contribution in this paper - namely, the multi-branch CNN architecture capable of processing the raw 1-d signal input (speech signal for SER) to create multiple spectrographic feature maps with a wide range of time-frequency resolution trade-offs. It can be seen that the input raw signal (shown as 1.5 sec duration here, made of 66150 samples corresponding to a sampling rate of 44.1 kHz), is fed to M branches, each with a set of 32 kernels, with each branch having a fixed kernel size (e.g. branch 1 has kernel size of 11 samples, branch 2 has kernel size 51 and so on). We consider in this work M up to 12, i.e. 12 branches, with $M = 12^{th}$ branch having the longest kernel of size 1501 samples.

To provide a reference, a conventional CNN has only one branch (with multiple kernels, e.g. 32 here), with some fixed size kernel size, e.g. 51 (in the 2nd branch). In such a conventional CNN branch, each kernel convolves with the 1-d signal input and yields an output that is a linearly filtered version of the signal through each of the 32 kernels in that branch. As the CNN learns to map the input to the classes in the fully connected layer in the output, the kernels (the filter coefficients) are optimized to learn to extract an appropriate feature signal from the input signal, and create a 'feature map' which is one spectrogram-like output made of 32 channels each with its time varying filter outputs. This 'single' spectrogram is governed by the time-frequency trade-off inherent and defined by the kernel size (of the single branch).

The resultant spectrogram-like feature map can be viewed as a narrow-band or wide-band spectrogram depending on the kernel size, as is well known for instance in speech signal processing [17], i.e., small kernels yielding high temporal resolution and poor frequency resolution resulting in a wide-band spectrogram and long kernels yielding poor temporal resolution and very good frequency resolution resulting in a narrow-band spectrogram. This can also be viewed as equivalent to a filter-bank analysis of the input signal with the filter-banks' filter's spectral characteristics (the band-pass bandwidths determined by the kernel size and the frequency response determined by the kernel values which in turn are determined by the CNN's weight learning for the given task).

It is clear that such a 'single' branch and the corresponding spectrogram with a time-frequency trade-off specific to the kernel size of that branch is highly restricted in the kind of time-frequency analysis it can perform on the input 1-d signal. For instance, in a wide class of 1-d signal classification problems such as speech recognition, audio-classification, music-genre classification problems or particularly the SER problem considered here, the signal is highly non-stationary with the spectral dynamics changing at varying rates in time, and with various spectral events localized in frequency likewise exhibiting different temporal evolutions. In order to capture these dynamic events in time and frequency, localized at different scales in time and frequency, a single spectrographic representation as obtained by a single branch CNN is clearly inadequate. This calls for a mechanism to generate time-frequency representations at different time-frequency resolutions, that is made possible by considering multiple branches in the CNN, with each branch with a pre-specified but variable kernel size which is same for all the kernels in that branch. Fig. 4a) shows such a multi-branch CNN in section marked 'A', with up to M branches. Shown are branches 1, 2 and 3 and $M = 12$, with the corresponding kernel sizes 11, 51, 101, 151,201, 251, 301, 501, 601, 751, 1001 and 1501. Such a multi-branch CNN will generate a spectrographic feature-map in 'each' of the M branches, each such feature map having its unique time-frequency trade-off determined by the kernel size used in the corresponding branch. For example, here, Branch 1 with kernel size 11 samples, will yield a very wide-band spectrogram (with a very fine time-resolution and poor frequency resolution), Branch 2 with kernel size 51 samples will yield a less wide-band spectrogram,

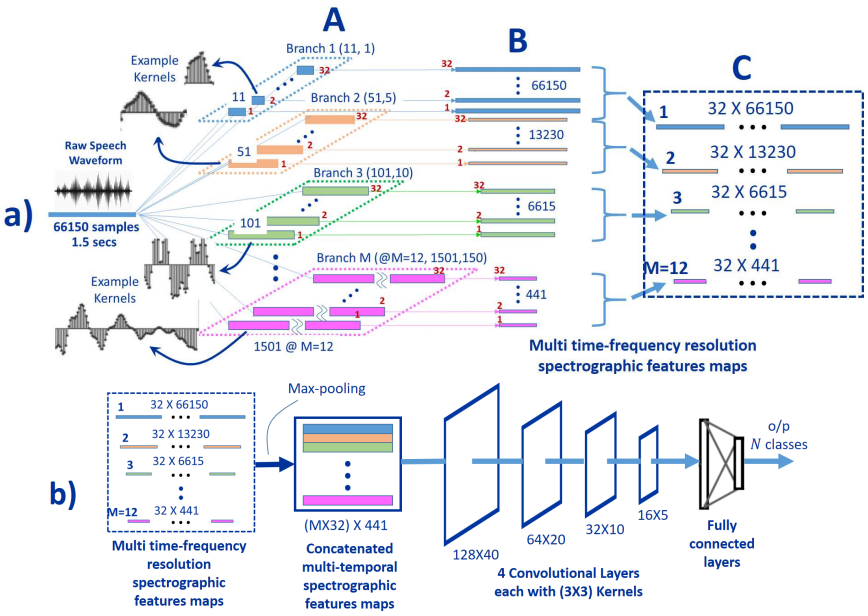

Figure 4: *Multi-temporal CNN architecture* - **a)** *formation of multi time-frequency spectrographic feature maps,* **b)** *from multi time-frequency spectrographic feature maps to fully connected layers*

Branch 3 with a kernel size 101 samples will yield a narrow-band spectrogram, and Branch $M = 12$ with a very long kernel size 1501 samples will yield a very narrow-band spectrogram (with a poor time-resolution and very good frequency resolution). Thus the M branches taken together will yield a multi-temporal time-frequency resolution spectrographic feature map (as shown in the sections marked 'B' and 'C' in this figure), each of size 32 frequency channels $\times$ number of filter outputs decided by the stride of the convolution kernel in that branch (e.g. 32×6615 for Branch 3 with stride of 100).

The feature maps in 'C' are a stack of 32 individual spectrographic maps, each of length (66150, 13250, 6615, . . ., 441) corresponding to the 12 branches, and each of these are subject to max-pooling to reduce them to a feature-map of size $(M \times 32)\times 441$ or 384×441 for $M = 12$. This is shown in Fig. 4b) outlined further below.

The feature map stack in 'C', on being reduced to a feature-map of size 384×441 for $M = 12$, as shown in Fig. 4(b) is further processed by 4 convolutional layers, each with 64, 128, 256 and 256 filters each filter being a 3×3 kernel with a stride 1×1, yielding respectively 64 (128×40), 128 (64×20), 256 (32×10) and 256 (16×5) feature maps on suitable max-pooling at each stage. The final output of size $256 \times 16 \times 5$ from the fourth convolution layer is used directly as input to the fully-connected layer with an output layer with N soft-max outputs (corresponding to N classes; $N = 5$ for the 5 emotional classes in IEMOCAP data-set chosen here). The feature map stack in 'C' (and forming the input of 384×441) represents the joint feature map across the multi-temporal multi time-frequency resolution spectrographic feature maps (multi time-frequency textures in the stack representing the input emotional speech from the raw speech waveform) and captures the different time-frequency event localizations that would be present in the input 1-d speech signal.

As a means of benchmarking the architecture's performance, prior to applying it to the OCD patient data, we show the SER accuracy of this architecture on the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database [18] for 5 classes namely, Anger, Frustation, Excited, Neutral and Sad (i.e., $N = 5$ in Fig. 4) each having 2460 sec of speech. We used a 70:30 (train, test) split and a 5-fold validation with input durations of 2 secs. Fig. 5 shows the accuracy of SER task for the i) single-temporal cases (single branches with kernel sizes 11, 51, 101, 151, 201, 251, 301, 501, 601, 751, 1001 and 1501) and ii) the performance gain of 2-8% (absolute) of the multi-branch architectures for different $M = 3, 6, 9, 12$ - termed 3-branch, 6-branch, 9-branch and 12-branch architecture with respect to the individual single-branch performances in (i). The results demonstrate the advantage of the multi-temporal architecture over a single temporal architecture (the conventional CNNs) - for which it is clear that the best kernel size has to be predetermined for a given task, such as the kernel sizes of 51, 101, 151 offering the best performance for the SER task here, but which is obviated by the use of a multi-temporal architecture with even as small as 3 or 6-branches.

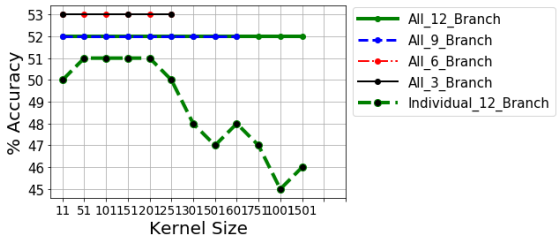

Figure 5: *Performance of proposed multi-temporal CNN architecture*

4. Position paper methodology

We propose through this position paper a methodology of adapting this architecture for arousal-valence estimation from the input raw speech signal of OCD patients in the ERP therapy, by using a loss function based on the concordance correlation coefficient (CCC) and trained on raw valence-arousal annotated data (such as in [19], [20]) and apply it to the OCD scenario. The valence-arousal estimates are further transformed into fine-granular emotional states and degree of emotion using the 3-d or 2-d model as outlined in Sec. 2.4.

5. References

[1] Bjorn W. Schuller. Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. Commuications of the ACM, vol. 61, no. 5, pp. 90-99, May 2018.

[2] Dianne M. Hezel and H. Blair Simpson. Exposure and response prevention for obsessive-compulsive disorder: A review and new directions. Indian J Psychiatry. 2019 Jan; 61(Suppl 1): S85S92.

[3] http://beyondocd.org/information-for-individuals/cognitive-behavior-therapy

[4] Jonathan S Abramowitz. The Psychological Treatment of Obsessive-Compulsive Disorder. Can J Psychiatry, Vol 51, No 7, pp. 407-416, June 2006.

[5] Maureen L. Whittal, Dana S. Thordarson and Peter D. McLean. Treatment of obsessivecompulsive disorder: Cognitive behavior therapy vs. exposure and response prevention. Behaviour Research and Therapy 43, pp. 15591576, 2005.

[6] Michael G. Wheaton, Hanga Galfalvy, Shari A. Steinman, Melanie M. Wall, Edna B. Foa, and H. Blair Simpson. Patient adherence and treatment outcome with exposure and response prevention for OCD: Which components of adherence matter and who becomes well?. Behav Res Ther., 85: 612, Oct 2016.

[7] https://medicine.umich.edu/sites/default/files/content/downloads/Exposure-and-Desensitization.pdf

[8] Roh D, Kim W. J., Kim C. H. Alexithymia in obsessive-compulsive disorder: clinical correlates and symptom dimensions. J Nerv Ment Dis., 199(9):690-5, Sep 2011.

[9] Andrea Pozza, Nicoletta Giaquinta and Davide Dttore. The Contribution of Alexithymia to Obsessive-Compulsive Disorder Symptoms Dimensions: An Investigation in a Large Community Sample in Italy. Psychiatry Journal, vol. 2015, Article ID 707850, 6 pages.

[10] https://www.anxietycanada.com/sites/default/files/adult_hmocd.pdf

[11] James A. Russell. A circumplex model of affect. Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 11611178, 1980.

[12] http://www.adliterate.com/archives/Plutchik.emotion.theorie.POSTER.pdf

[13] Dorota Kaminska and Tomasz Sapinski and Adam Pelikant. Recognition of emotion intensity basing on neutral speech model. Man-Machine Interactions 3 (pp.451-459), 2014. Springer (eds): A.Gruca, T.Czachrski, S.Kozielski.

[14] Carroll Ellis Izard. The face of emotion. Century psychology series, Appleton-Century-Crofts, 1971 (Digitized 2010).

[15] Sven Buechel and Udo Hahn. Readers vs. Writers vs. Texts: Coping with Different Perspectives of Text Understanding in Emotion Annotation. https://www.researchgate.net/publication/316546282

[16] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan. Combining categorical and primitives-based emotion recognition. Proc. EUSIPCO 2006.

[17] T. F. Quatieri. Discrete Time Speech Signal Processing. Prentice Hall, 2002

[18] https://sail.usc.edu/iemocap/

[19] Panagiotis Tzirakis, Jiehao Zhang, Bjorn W. Schuller. End-to-end speech emotion recognition using deep neural networks. Proc. ICASSP '18, pp. 5089-5093, 2018.

[20] N. Malandrakis, A. Potamianos, G. Evangelopoulos and A. Zlatintsi. A supervised approach to movie emotion tracking. Proc. ICASSP '11, 2011.