# Dimensional Emotion Recognition from Speech Using Modulation Spectral Features and Recurrent Neural Networks

Zhichao Peng[*, †], Zhi Zhu[§], Masashi Unoki[*], Jianwu Dang[*, †], Masato Akagi[*]
[*]School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[†]School of Computer Science and Technology, Tianjin University, Tianjin, China
[§]Fairy Devices Inc., Tokyo, Japan
E-mail: {zcpeng, zhuzhi, unoki, jdang, akagi} @jaist.ac.jp

***Abstract*—Dimensional emotion recognition (DER) from speech is used to track the dynamics of emotions for robots to naturally interact with humans. The DER system needs to obtain frame-level feature sequences by selecting the appropriate acoustic features and duration. Moreover, these sequences should reflect the dynamic characteristics of the utterance. Temporal modulation cues are good at capturing the dynamic characteristics for speech perception and understanding. In this paper, we propose a DER system using modulation spectral features (MSFs) and recurrent neural networks (RNNs). The MSFs are obtained from temporal modulation cues, which are produced from auditory front-ends by auditory filtering of speech signals and modulation filtering of the temporal envelope in a cascade manner. Then, the MSFs are fed into RNNs to capture the dynamic change of emotions from the sequences. Our experiments of predicting valence and arousal involving the RECOLA database demonstrated that the proposed system significantly outperforms the baseline systems, improving arousal predictions by 17% and valence predictions by 29.5%.**

## I. INTRODUCTION

Dimensional emotion describes more mixed emotions and captures the gradual emotion transitions in spontaneous or natural speech [1]. In human-robot interaction (HRI), robots need to capture the continuous changes of the speaker's emotions in order to interact naturally. Therefore, dimensional emotion can better meet the needs of HRI than discrete emotion can. Researchers have gradually shown an increasing interest in the representation and recognition of dimensional emotions [2]. Valence and arousal are the universal primitives in emotion dimensional space. Valence is related to the subjective appraisal and experience of positive or negative emotion. Arousal is associated with an intensity level, particularly strong or weak. Dimensional emotion recognition (DER) is mainly studied from two aspects. One is how to select the appropriate acoustic features and duration to extract frame-level feature sequences that can reflect the dynamic characteristics of the utterance. The other is how to capture the dynamic changes in emotional states from feature sequences.

To extract and select emotional features, most existing DER systems extract acoustic features from sequential low-level descriptors (LLDs), such as F0 and Mel-frequency cepstral coefficients (MFCCs), to match the granularity of the annotation in each primitive. The values of each primitive are continuously labeled on short-time frames, such as 40ms in the RECOLA database [3]. MFCCs are the most commonly used feature in speech emotion, but they cannot reflect the dynamic characteristics of speech signals very well. Previous studies have indicated that temporal modulation cues are good at capturing the temporal dynamic cues for speech perception and understanding [4, 5]. Moreover, Dau et al. [6] implemented auditory perception models to simulate the results from different psychoacoustical tests. Additionally, some studies extracted temporal modulation cues on the basis of the auditory perception models for emotional speech analysis and showed that the cues are important for emotion recognition [7-10]. Wu et al. [9] proved that the modulation spectral features (MSFs) consistently exhibit considerably better discrimination power than MFCCs for categorical emotion recognition. In our previous studies, we proved modulation spectral representations can extract salient emotion features on spectral-temporal representations for two-stage [10] and end-to-end categorical emotion recognition [11] using various convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

Next, a regression model should be considered to capture the dynamic changes of emotional states from feature sequences in dimensional emotion recognition. Long Short-Term Memory (LSTM) recurrent networks can capture the temporal information to predict continuous dimensional values and explore many variations to improve performance. Some studies have used LSTM to predict dimensional emotions [12–14]. Wöllmer et al. [12] presented a fully automatic audiovisual recognition approach based on LSTM modeling of word-level audio and visual features. LSTM achieved a higher prediction accuracy than Support Vector Regression (SVR) due to its ability to model long-term time dependencies and decrease the time delay. Trigeorgis et al. [14] used a 1D convolutional operation directly on the discrete-time speech signals to predict dimensional emotions. In this paper, we propose a DER system to predict valence and arousal primitives using auditory-inspired MSFs and LSTM. Figure 1 shows an overview of the proposed DER system.

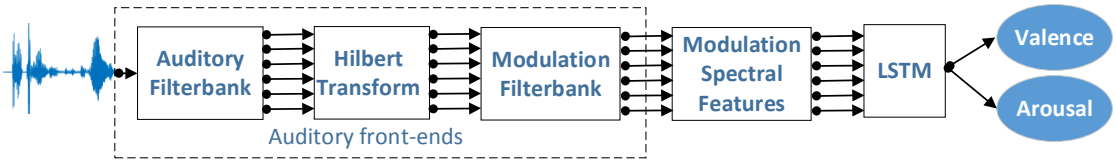


Fig. 1 Overview of proposed DER system

The modulation spectral representation is obtained from auditory front-ends including auditory filtering of speech signals and modulation filtering of the Hilbert temporal envelope in a cascade manner. Then the MSFs are extracted from modulation spectral representations by computing the frame-level centroid, skewness, kurtosis, flatness, tilt, etc. Eventually, LSTM is used to model the temporal-dynamics information for continuous dimensional emotion recognition. The experiments of predicting valence and arousal are conducted in the RECOLA database. We also investigate the effects of MSFs with different window lengths on predicting valence and arousal.

The paper is organized as follows. Section 2 introduces the extraction of MSFs from auditory front-ends. Section 3 presents the proposed regression model for DER. Section 4 describes and discusses our experimental results. Finally, section 5 concludes the paper.

## II. Auditory-inspired Modulation Spectral Features

In this section, we describe the auditory front-ends model to produce temporal modulation cues, and then different MSFs are extracted from the modulation spectral representation in the acoustic and modulation frequency domains.

### *A. Temporal modulation cues from auditory front-ends*

In the auditory front-end, the emotional speech signal $s(t)$ is first filtered by a bank of auditory filters to emulate the processing carried out by the cochlea.

The output of the ith-channel signal is given by

$$s(i,t) = gt(i,t) * s(t),\ 1 \le i \le N, \quad (1)$$

where $gt(i,t)$ is the impulse response of the ith channel, t is the sample number in the time domain, and N is the number of auditory filters. The center frequencies of these filters are proportional to their bandwidths, which in turn are characterized by the equivalent rectangular bandwidth (ERB) [15]:

$$ERB_i = \frac{f_i}{Q_{ear}} + B_{min}, \quad (2)$$

where $f_i$ is the center frequency (in Hz) of the ith critical-band filter, and $Q_{ear}$ and $B_{min}$ are constants set to 9.26449 and 24.7, respectively. The impulse response of a Gammatone filter is the product of a Gamma distribution and a sinusoidal ton [16].

$$gt(i,t) = At^{n-1}exp(-2\pi b_f ERB(f_i)t)\,cos(2\pi f_i t), \quad (3)$$

where $At^{n-1}\exp(-2\pi b_f \mathrm{ERB}(f_i)t)$ is the amplitude term represented by the Gamma distribution, A, n, *and* $b_f$ are the amplitude, filter order, and bandwidth of the filter, respectively.

This envelope is extracted using the Hilbert transform to calculate the instantaneous amplitude $H(i,t)$ of the ith channel signal. The $H(i,t)$ is computed from $s(i,t)$ as the magnitude of the complex analytic signal $\hat{s}(i,t) = s(i,t) + \mathrm{jH}\{s(i,t)\}$, where H{·} denotes the Hilbert transform. Hence,

$$H(i,t) = |\hat{s}(i,t)| = \sqrt{s^2(i,t) + H^2\{s(i,t)\}}. \quad (4)$$

Furthermore, the modulation filterbank of M sub-channels is used for the kth sub-channel in the ith channel signal to extract the spectral-temporal representation signal $MSR(i,k,t), 1 \le k \le M$. Since the values of each primitive are continuously labeled on short-time frames, we extracted short-term MSFs as a frame to match values of each primitive. For each frame x, the modulation spectral representation is represented as $\mathrm{MSR_x(i,k)}$.

Figure 2 shows the 2D spectrogram of modulation spectral representations with the first three channels (one low-pass filter and the first two band-pass filters) for the first 30 seconds of the utterance "P58" in the RECOLA database. The top panel is the spectrogram referred to the low-pass filter with cut-off frequency $\mathrm{f_{cut}} = 2\mathrm{Hz}$ in the modulation frequency channel. The middle and bottom panels are the spectrogram referred to the band-pass filter with frequencies of 4 and 8 Hz, respectively. The modulation spectral representations are composed of 288 channels with 32 acoustic channels and 9 modulation channels in this study. The higher energy is mostly concentrated at the lower-modulation-frequency channel. This means that the lower modulation frequency plays a greater role in emotion recognition. However, high-modulation-frequency channels may contain information such as fundamental frequency and may help to estimate the values of valence and arousal primitives.

### *B. Modulation spectral feature extraction*

We extracted 13 types of MSFs to determine whether these features can be used to identify the dimensional emotion. Two kinds of MSFs were calculated by analyzing the modulation spectral representation in the acoustic and modulation frequency domains. In the acoustic frequency domain, the first feature is the mean of the modulation spectrum $\mathrm{MSR_x(i,k)}$ in the kth modulation channel and represents the energy distribution of the modulation frequency, specifically:

$$\varphi_{ma}(k) = \frac{\sum_{i=1}^{N} MSR_x(i,k)}{N}. \quad (5)$$

The second feature is the modulation spectral centroid $\varphi_{\mathrm{ca}}(\mathrm{k})$, which indicates the center of the spectral balance across acoustic frequency bands. $\varphi_{\mathrm{ca}}(\mathrm{k})$ is defined as:

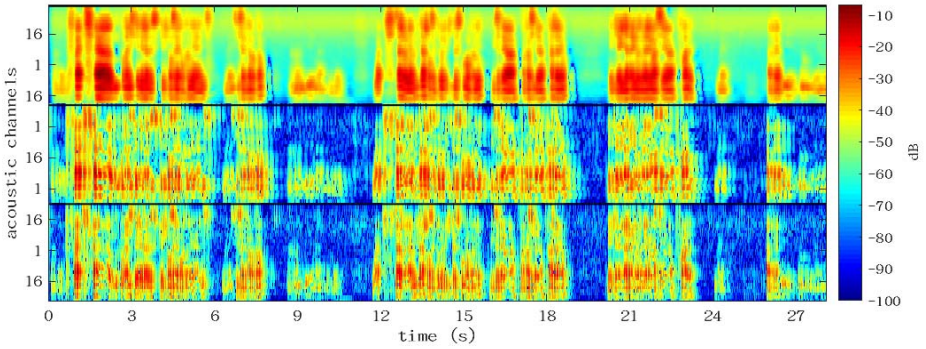


Fig.2 2D spectrogram of modulation spectral representations

$$\varphi_{ca}(k) = \frac{\sum_{i=1}^{N} iMSR_x(i,k)}{\sum_{i=1}^{N} MSR_x(i,k)}. \quad (6)$$

The third feature is the modulation spectral spread $\varphi_{sa}(k)$, which represents the spread of the spectrum around its spectral centroid as the second moment. $\varphi_{sa}(k)$ is defined as:

$$\varphi_{sa}(k) = \frac{\sum_{i=1}^{N} [i-\varphi_{ca}(k)]^2 MSR_x(i,k)}{\sum_{i=1}^{N} MSR_x(i,k)}. \quad (7)$$

The fourth feature is modulation spectral skewness $\varphi_{ska}(k)$, which describes the degree of asymmetry of the spectrum around its spectral centroid as the third moment. $\varphi_{ska}(k)$ is defined as:

$$\varphi_{ska}(k) = \frac{\sum_{i=1}^{N} [j-\varphi_{ca}(k)]^3 MSR_x(i,k)}{\sum_{i=1}^{N} MSR_x(i,k)}. \quad (8)$$

The fifth feature is modulation spectral kurtosis $\varphi_{kta}(k)$, which describes the measurement of the peakedness of the spectrum around its spectral centroid as the fourth moment. $\varphi_{kta}(k)$ is defined as:

$$\varphi_{kta}(k) = \frac{\sum_{i=1}^{N} [j-\varphi_{ca}(k)]^4 MSR_x(i,k)}{\sum_{i=1}^{N} MSR_x(i,k)}. \quad (9)$$

The sixth feature is modulation spectral flatness $\varphi_{sfa}(k)$, which is computed from the ratio of the geometric mean of the arithmetic mean of the spectrum. $\varphi_{sfa}(k)$ is defined as:

$$\varphi_{sfa}(k) = \frac{\sqrt[N]{\prod_{i=1}^{N} iMSR_x(i,k)}}{\varphi_{ma}(k)}. \quad (10)$$

The seventh feature is modulation spectral tilt $\varphi_{sta}(k)$, which represents the linear regression coefficient obtained by fitting a first-degree polynomial to the modulation spectrum in dB scale. $\varphi_{sta}(k)$ is defined as:

$$\varphi_{sta}(k) = \frac{\sum_{i=1}^{N} iMSR_x(i,k)}{\sum_{i=1}^{N} MSR_x(i,k)}. \quad (11)$$

On the modulation frequency domain, the six other features are calculated in the ith acoustic channel. For example, the spectral centroid $\varphi_{cm}(i)$ in the acoustic frequency domain is computed as follows:

$$\varphi_{cm}(i) = \frac{\sum_{k=1}^{M} kMSR_x(i,k)}{\sum_{k=1}^{M} MSR_x(i,k)}. \quad (12)$$

Then we obtain the mean of energy, spectral spread, spectral skewness, spectral kurtosis, and spectral tilt using similar formula on the modulation frequency domain. Lastly, we obtain 63 acoustic-frequency-domain features and 192 modulation-frequency-domain features, thus totaling 255 features.

## III. EMOTION RECOGNITION MODELS

In this section, we introduce the emotion recognition models for DER. The temporal RNN model and the multi-task learning framework are described below.

### *A. RNN model*

LSTM architecture is the state-of-art model for sequence analysis since it can exploit long-term dependencies in the sequences by using memory cells to store information. Given an input feature sequence $x = \{x_1, \dots, x_T\}$, LSTM computes the hidden vector sequence $h = \{h_1, \dots, h_T\}$ and output vector sequence $y = \{y_1, \dots, y_T\}$ by iterating the following equations from $t = 1$ to T:

$$(h_t, c_t) = H(x_t, h_{t-1}, c_{t-1}), \quad (13)$$

$$y_t = w_y * h_t + b_y, \quad (14)$$

where the H term is the LSTM layer function, c is the cell activation vector with the same size as the hidden vector h. The w and b terms denote the weight matrices and bias vectors, respectively. The LSTM layer is composed of one LSTM cell, a dropoutwrapper with keep probability of 0.5, and peephole connections.

### *B. Performance measure and loss function*

To measure the weight of each feature, a concordance correlation coefficient (CCC) measure between the prediction values of emotion dimensions and the gold standard values is used. $\rho_c$ is a measure of how well the prediction values of emotion dimensions (Y) compares to a "gold standard" measurement (X).

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}, \quad (15)$$

where ρ is the Pearson correlation coefficient (PCC) between the two time series prediction and gold-standard, $\sigma_x^2$ and $\sigma_y^2$ is

the variance of each time series, $\mu_x$ and $\mu_y$ are the mean values of each. Therefore, predictions that are well correlated with the gold standard but shifted in value are penalized in proportion to the deviation. This means that the CCC measure combines the PCC with the square difference between the mean of the two compared time series. Hence, we utilize a loss function ($L_c$) on the basis of the concordance correlation coefficient. $L_c$ is defined as:

$$L_c = \frac{2-\rho_c^a-\rho_c^v}{2}, \quad (16)$$

where $\rho_c^a$ and $\rho_c^v$ are the CCC of the arousal and valence, respectively.

### C. Multitask learning

As the arousal and valence are highly correlated with each other, we propose multi-task learning to predict the arousal and valence simultaneously in the DER system. We train a LSTM regression model with two outputs and two CCC losses at the same time.

## IV. Experimental Results

### A. Database Description

The RECOLA database is a multi-modal corpus of remote collaborative and affective interaction in French. The version used in this study contains 23 conversations, each lasting 5 minutes. To ensure speaker-independence in the experiments, the corpus was split into three partitions, by balancing the gender and the age of the subjects: training (9 subjects), validation (9 subjects), and testing (5 subjects). Annotation was performed for arousal and valence separately. Affective behavior of the participants was evaluated by six different annotators and averaged over all annotator by considering inter- annotator agreement to provide a gold standard.

### B. Experimental Setup

For preprocessing, we normalize MSFs on the feature level to reduce the difference between different features. Due to the high time-resolution of the modulation spectral representations, the number of samples for the time domain has to be reduced. The time-resolution is reduced simply by downsampling modulation spectral representations with an 800-Hz rate.

The structure of LSTM contains an input layer with 255-dimensional input features, 1 hidden layer with 256 hidden units, and then following a full-connected layer with 128 outputs, an output layer with 2 nodes corresponding to the predicted valence and arousal primitives. Additionally, for all random weight initializations, we choose L2-regularliser initialization. To counter overfitting in training our regression model, we use a dropout strategy before the fully-connected layer with a dropout rate of 0.5. We trained our model throughout all experiments with the Adam optimizer with a fixed learning rate of 1e-4. The mini-batch size utilized was 10 with a sequence length of 1500 frames (60 s) when training, and the model is tested on the entire records without segmentation. We trained the regression model with a maximum 200 epochs and stopped training when the predictions did not improve for either dimension on the validation set for 10 epochs.

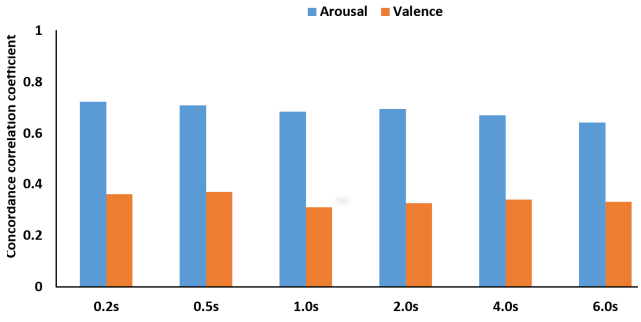


Fig. 3 Results comparing the prediction of MSFs under different windows length

### C. Prediction of MSFs under different window lengths

We investigate the effects of different window lengths on predicting valence and arousal. Different window lengths ranging from 0.2-6s were explored and 255 MSFs are computed per window. Figure 3 depicts the obtained CCC for arousal and valence. Results show that the best performance for arousal is achieved with a 200-ms window, resulting in a CCC of 0.724. For valence, the best performance for arousal is achieved with a 500-ms window, resulting in a CCC of 0.37. This means that the MSFs with a 200-ms window contain temporal dynamic cues. Hence, we select the window length of 200 ms to extract features for emotion recognition in this study.

### D. Comparison with different methods

In this work, we first implement a baseline model with MFCCs and a SVR model. To extract acoustic features from the speech recordings, we used the openSMILE toolkit [17] to extract 39 MFCCs (12 MFCC + energy, 12 delta MFCC + energy and 12 double delta MFCC + energy) with a frame window size of 25 ms at a step of 10 ms. We stack four frames to form a 40-ms feature vector, thus totaling 156 MFCC features. Then, a SVR predictor is used to recognize emotion.

In addition, we implement our temporal model and multi-task strategy with the tensorflow deep learning framework. We compare our approach with other emotion recognition approaches.

Results obtained for each method are shown in Table 1. The highest performance was achieved by applying MSFs and LSTM. Arousal predictions improved by 17%, going from 0.619 to 0.724, while valence improved 29.5%, going from 0.278 to 0.36. Yang and Hirschberg [18] obtained a CCC with

Table 1. Performance comparison (in term of $\rho_c$) under different features and predictors

| | | Arousal | | Valence | |
|---|---|---|---|---|---|
| Predictor | Features | Dev | Test | Dev | Test |
| SVR | MFCC | .657 | .619 | .320 | .278 |
| LSTM | MFCC | .696 | .674 | .346 | .294 |
| SVR | MSFs | .713 | .683 | .342 | .287 |
| **LSTM** | **MSFs** | **.751** | **.724** | **.367** | **.36** |

0.692 and 0.423 using the same dataset after training deep neural networks on waveforms and spectrograms. In contrast, our system obtained a better CCC in arousal prediction.

### *E. Discussion*

In our preliminary experiments, we further found that bidirectional LSTM and multi-layer LSTM do not improve the performance. Similar to the results of Trigeorgis et al. [14], no feature delay is required to compensate for the cognition delay of the annotators for MSFs since the LSTM networks can capture the temporal information and long-term time dependencies. However, the centering by finding the ground truth's and the prediction's biases is useful to improve the performance.

## V. CONCLUSIONS

In this study, we proposed a dimensional emotion recognition system to predict valence and arousal primitives using 255-dimension MSFs and LSTM recurrent networks. We showed that MSFs can effectively extract frame-level emotional features from continuous speech signals and LSTM can capture the dynamic changes of emotional states from feature sequences. Future work will include robustly analyzing the proposed system in a noisy environment and further investigating the 3D modulation spectral representations (acoustic frequency components, modulation frequency components, and temporal features) for emotion recognition.

## ACKNOWLEDGEMENTS

This study was supported by the Research Foundation of Education Bureau of Hunan Province, China (Grant No. 18A414)

## REFERENCES

[1] G. Valenza, S. Member, and A. Lanata, "The Role of Nonlinear Dynamics in Affective Valence and Arousal Recognition," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 237–249, 2012.

[2] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. Di Natale, "Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 314–327, 2017.

[3] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," *2013 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013*, no. i, 2013.

[4] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Appl. Signal Processing*, vol. 2003, no. 7, pp. 668–675, 2003.

[5] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 585–592, 1995.

[6] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2906–2919, 1997.

[7] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Modulation Spectral Features for Predicting Vocal Emotion Recognition by Simulated Cochlear Implants," *Interspeech 2016*, pp. 262–266, 2016.

[8] A. R. Avila, Z. A. Momin, J. F. Santos, D. O'Shaughnessy, and T. H. Falk, "Feature Pooling of Modulation Spectrum Features for Improved Speech Emotion Recognition in the wild," *IEEE Trans. Affect. Comput.*, vol. 3045, no. c, pp. 1–1, 2018.

[9] S. Wu, T. H. Falk, and W. Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011.

[10] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech Emotion Recognition Using Multichannel Parallel Convolutional Recurrent Neural Networks based on Gammatone Auditory Filterbank," *Apsipa Asc*, pp. 0–5, 2017.

[11] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Auditory-inspired end-to-end speech emotion recognition using 3d convolutional recurrent neural networks based on spectral-temporal representation," *2018 IEEE Int. Conf. Multimed. Expo*, pp. 1–6, 2018.

[12] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput.*, vol. 31, no. 2, pp. 153–163, 2013.

[13] S. Chen and Q. Jin, "Multi-modal Dimensional Emotion Recognition Using Recurrent Neural Networks," *Proc. 5th Int. Work. Audio/Visual Emot. Chall.*, pp. 49–56, 2015.

[14] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," *Proc. 41st IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 5200–5204, 2016.

[15] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753, 1983.

[16] R. D. Patterson, M. Unoki, and T. Irino, "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *J. Acoust. Soc. Am.*, vol. 114, no. 3, pp. 1529–1542, 2003.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM int. conf. Multimedia*, 2010, pp. 1459–1462.

[18] Z. Yang and J. Hirschberg, "Predicting Arousal and Valence from Waveforms and Spectrograms using Deep Neural Networks," *Proc. Interspeech 2018*, pp. 3092–3096, 2018.