SPEECH-BASED STRESS CLASSIFICATION BASED ON MODULATION SPECTRAL FEATURES AND CONVOLUTIONAL NEURAL NETWORKS

Anderson R. Avila[1], Shruti R. Kshirsagar[1], Abhishek Tiwari[1], Daniel Lafond[2], Douglas O'Shaughnessy[1], and Tiago H. Falk[1]

[1]Institut national de la recherche scientifique (INRS-EMT), Quebec, Canada
[2]Thales Research and Technology, Canada, Quebec, Canada

ABSTRACT

Interest in stress recognition has notably increased over the past few years. In this work, we focus on recognizing stress from speech. We propose the use of modulation spectral features as input to a convolutional neural network (CNN) for classifying stress. As benchmark, the OpenSMILE features used in the INTERSPEECH 2010 Paralinguistic Challenge is adopted and evaluated with a support vector machine (SVM) and a deep neural network (DNN) based backends. Experiments are performed with the well-known Speech Under Simulated and Actual Stress (SUSAS) database. Performances are investigated considering 2-class, 4-class and 9-class classification problems. Results show that the proposed approach outperforms the benchmark on a challenging 9-class classification task with accuracy as high as 70% representing gains of roughly 18% over the benchmark.

Index Terms— Stress detection, modulation spectrum, convolutional neural network

1. INTRODUCTION

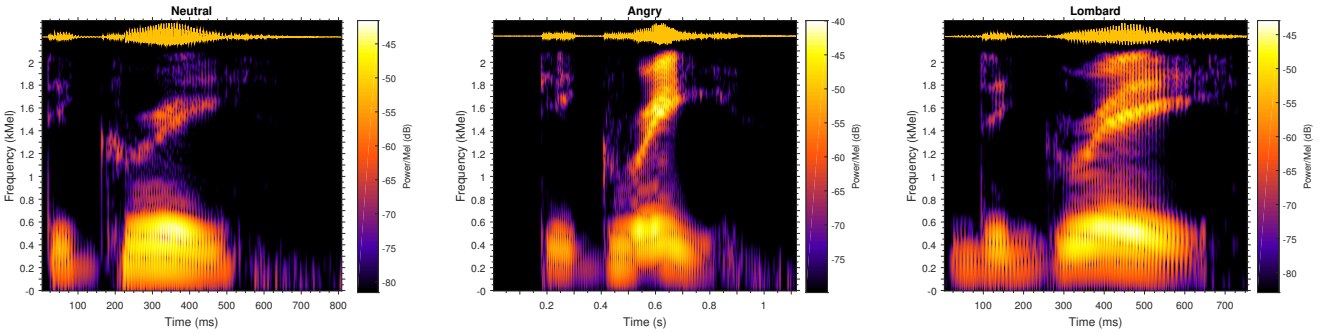
Mental stress has become a recurrent threat to modern society. If not treated in early stages, it can become a chronic condition leading to serious health problems [1, 2]. Its detrimental effects, for instance, can range from physical (e.g., cardiovascular disease) to psychological such as depression and sleep disorders [3, 4]. A large component of stress is linked to the workplace, with as much as 50% of employees suffering from "work stress" [5]. This type of stress has often been associated with job performance degradation [6]. Moreover, for critical jobs, such as first responders and air traffic controllers, to name a few, stress can lead to drastic consequences. Post traumatic stress disorders can even lead to suicidality [7].

Stress influences the human autonomic nervous system (ANS), altering the heart rate, breathing rate, fatigue levels, as well as the muscle tension of the vocal chord, specially while performing a secondary physical task [1]. As such, stress can impact the way we produce speech [8]. Figure 1, for example, depicts three spectrograms for neutral, angry and speech produced under the lombard effect. As can be seen, for angry and lombard speech, the spectrogram shows more energy in higher frequencies when compared to the neutral speech produced by the same speaker. Note that shift of the fundamental frequency $F0$, as well as the presence of more prominent formants and higher average frequencies are normally expected with increased levels of stress or type of emotion [9].

Although spectrum variability may benefit stress detection from speech, multi-class stress recognition is still a challenging task. Several features such as pitch, energy, spectral band energy, and cepstral coefficients have been explored for stress and emotion detection. The authors in [10], for instance, proposed a set of high order features for emotion/stress detection using support vector machine (SVM) and extreme learning machine (ELM) as backend. Such features were combined with the so-called Interspeech 2010 features to improve recognition performance. In [11], in turn, perceptual content of voice quality, first- and second-order differences based on a new Fourier parameter (FP) model were proposed for speaker-independent speech emotion recognition. An improvement of 16.2% was achieved when combining the FP model with mel-frequency cepstral coefficients (MFCC). Recently, researchers have explored the use of deep neural networks (DNN) for emotion detection. In [12], an end-to-end multimodal system was proposed to recognize spontaneous emotion from raw speech and visual data. Although a number of studies have addressed the use of DNNs for emotion recognition, to the best of our knowledge, work has yet to emerge on the use of DNNs for speech-based stress detection.

In this paper, we aim to fill this gap. Recently, we proposed a new set of modulation spectral features (MSF) that, when combined with statistical pooling, resulted in accurate continuous speech emotion recognition "in-the-wild" [13, 14]. Notwithstanding, for short utterances (i.e. less than 3 seconds) the statistical pooling is expected to provide low performance as it depends on the analysis window length, typically 1 to 4 s. Hence, our goal is two-fold. First, we investigate if the MSF features can be applied to stress detection. Next, as the dataset used in our experiments contains short utterances, we propose the use of a convolutional neural net-

978-9-0827-9703-9/19/$31.00 ©2019 IEEE



(a) (b) (c)

Fig. 1: Spectrogram of speech in (a) neutral, (b) angry and (c) lombard for the same sentence and speaker.

work (CNN) in lieu of the previous statistical pooling to boost performance. Experiments are performed using the Speech Under Simulated and Actual Stress (SUSAS) database [15]. Results are compared to two benchmark systems based on features extracted from the OpenSMILE toolkit and a support vector machine (SVM) as well as a deep neural network (DNN) classifiers.

The remainder of this paper is organized as follows: Section 2 describes the proposed features and model. Section 3 gives details on the experimental setup and results are reported in Section 4. Conclusions are drawn in Section 5.

2. BACKGROUND: FEATURES AND MODELS

In this section, we describe the proposed features and the proposed convolutional neural network model.

2.1. Modulation Spectrum Features

The use of modulation spectral features (MSF) is motivated by our recent work [16, 14, 13], where MSFs were successfully employed in spontaneous speech emotion recognition in-the-wild. In these previous works, however, the focus was in continuous recognition of emotion primitives, such as valence, arousal, and dominance. For such task, feature pooling was shown to improve performance in realistic settings and statistical functionals were used [13]. These features have yet to be tested for emotion *classification* and for stress detection.

In order to extract the modulation spectral representation, the speech signal activity level is first normalized to -26 dBov (dB overload), eliminating unwanted energy variations among speech samples. Next, we filter the speech signal $\hat{x}(n)$ with a 23-channel gammatone filterbank, which simulates the cochlear processing [17]. The first filter of the filterbank is centered at 125 Hz and the last one at half of the sampling rate [13]. Each filter bandwidth follows the equivalent rectangular bandwidth (ERB), as described in [17]. The temporal envelope $e_j(n)$ is then computed using:

$$e_j(n) = \sqrt{\hat{x}_j(n)^2 + \mathcal{H}\{\hat{x}_j(n)\}^2}, \quad (1)$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform and $\hat{x}_j(n)$ is the output of the j-th acoustic filter. Temporal envelopes $e_j(n), j = 1, \ldots, 23$ are then windowed with a 256-ms Hamming window and shifts of 40 ms. The modulation spectrum, $E_j(m, f_m)$, is obtained after computing the discrete Fourier transform $\mathcal{F}\{\cdot\}$ of the temporal envelope $e_j(m; n)$

$$E_j(m; f_m) = |\mathcal{F}(e_j(m; n))|, \quad (2)$$

where m represents the m-th frame obtained after every Hamming window multiplication and f_m designates modulation frequency. The time variable n is dropped for convenience.

Lastly, an auditory-inspired modulation filterbank is used to group modulation frequencies into eight bands, motivated by evidence of similar modulation filterbank structure in the human auditory system [18]. The result of this computation is denoted as $\mathcal{E}_{j,k}(m)$, $k = 1, ..., 8$, where j indexes the gammatone filter and k the modulation filter. The filter center frequencies are equally spaced in the logarithmic scale from 4 to 128 Hz. From this representation, five feature sets can be extracted, as summarized in Table 1.

The first configuration $\mathcal{E}_1$ corresponds to the vectorization of the average 23×8 modulation spectrum energy matrix (i.e., the $\mathcal{E}_{j,k}$, averaged over all the m frames), thus corresponding to utterance level features. The second configuration, $\mathcal{E}_2$, is attained by appending $\mathcal{E}_1$ with 39 additional features, corresponding to average energy, spectral flatness, spectral centroid, slope, and root mean squared-error across grouped modulation bands. More details about these features can be found in [13]. The third configuration ($\mathcal{E}_3$), in turn, is the result of applying statistical feature pooling to $\mathcal{E}_1$ and combining it with $\mathcal{E}_1$. As in [13], pooling is performed using eight functionals, namely: mean, standard deviation, variance, kurtosis, skewness, range, min, and max. Similarly, the fourth configuration ($\mathcal{E}_4$) corresponds to the same

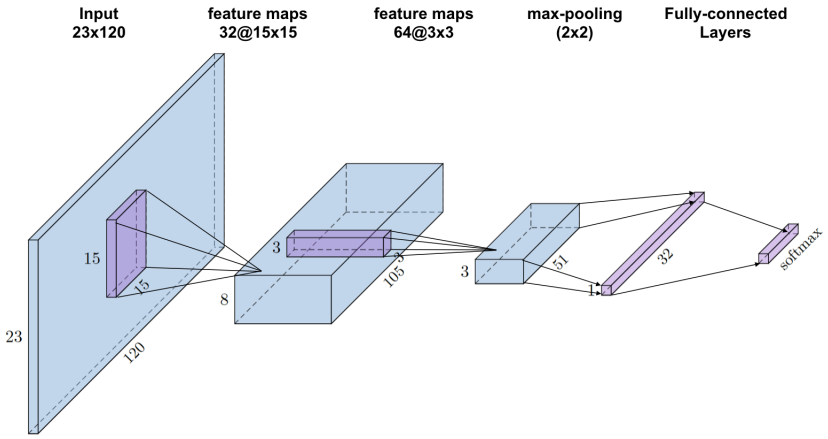


Fig. 2: Proposed CNN architecture.

Table 1: Dimensionality of MSF feature configurations

Configuration	Feature Dimensionality	After PCA
$\mathcal{E}_1$	184	140
$\mathcal{E}_2$	223	140
$\mathcal{E}_3$	1656	140
$\mathcal{E}_4$	1968	140
$\mathcal{E}_5$	23x120	N/A

8-functionals pooling, but of the $\mathcal{E}_2$ features. Principal component analysis (PCA) is also applied to reduce dimensionality to a set of more uncorrelated features, as depicted in Table 1. Final dimensions were defined empirically.

Lastly, the fifth configuration ($\mathcal{E}_5$) considers the $23 \times 8 \times M$ tensor configuration over all M modulation frames of a speech signal, as we wish to explore the use of a CNN to replace the functional pooling. To assure the CNN has inputs always of the same size, we propose the concatenation of the eight modulation bands for B consecutive active-speech modulation bands. Here, B is driven by the duration of the available speech data. For the SUSAS database, $B = 15$, thus resulting in a 23×120 input. In case of very short duration files, the last frame was replicated the number of times necessary to complete 120. Such a configuration was chosen empirically based on the average duration of the speech files.

2.2. Proposed CNN Architecture

Convolutional neural networks have successfully been applied to 2D image classification, as well as video, speech and audio processing applications [19]. CNNs are typically comprised of two layers: convolution and pooling. Convolutional layers are in charge of mapping, into their units, detected features from local connections in previous layers. Known as feature maps, this is the result of a weighted sum of the input features (or feature maps from previous convolutional layers) passed through a non-linearity such as ReLU [19]. A pooling layer will typically take the maximum or average of a set of neighboring feature maps, reducing dimensionality (i.e., subsampling) by merging semantically similar features.

As mentioned previously, here we are interested in exploring if CNNs can be used to replace the functionals-based pooling of [13] for emotional and stress classification. Our intuition is that a data-driven pooling mechanism can outperform functionals-based pooling and adapt to similar tasks, such as stress and emotion classification. The architecture proposed herein is depicted by Fig. 2. As can be seen, the model received as input a 23×120 matrix (stress detection from SUSAS), as described in Section 2.1. The first convolutional layer then applies 32 filters of 15×15 receptive field dimensionality. Each unit in the next layer is attained after applying ReLU as an activation function on each feature map. Another convolutional layer of 64 filters of 3×3 receptive field is then applied on the output of the previous layer, followed again by the ReLU activation function. Next, a max-pooling operation (2×2) is applied prior to a dropout (0.25) regularizer. A fully-connected layer is then used, followed by a dropout (0.25) and the softmax output unit. As optimizer, adadelta [20] was used as an adaptative leaning rate method.

3. EXPERIMENTAL SETUP

In this section, we describe the databases used, benchmark algorithm, classification tasks, and figure-of-merit.

3.1. Datasets and Dataset Partitioning

The SUSAS dataset is adopted in our experiments. The database is comprised of actual and simulated stress conditions. Only simulated stress was considered herein, as the actual settings involved helicopter pilots in action or riders in a roller coaster, thus the high levels of ambient noise could interfere with these initial tests. The SUSAS database was recorded from 32 speakers (13 female, 19 male) ranging

Table 2: Stress recognition accuracy (SUSAS dataset).

Classifier	Features	2 Classes	4 Classes	9 Classes	Average
SVM	OpenSMILE	68 %	57 %	58 %	61 %
	$\mathcal{E}_1$	61 %	53 %	44 %	52 %
	$\mathcal{E}_2$	63 %	53 %	46 %	54 %
	$\mathcal{E}_3$	63 %	54 %	49 %	56 %
	$\mathcal{E}_4$	63%	53 %	50 %	54 %
DNN	OpenSMILE	**83 %**	**77 %**	58 %	**72 %**
	$\mathcal{E}_1$	67 %	62 %	55 %	52 %
	$\mathcal{E}_2$	69 %	63 %	58 %	61 %
	$\mathcal{E}_3$	74 %	67 %	57 %	63 %
	$\mathcal{E}_4$	75 %	67 %	62 %	68 %
CNN	$\mathcal{E}_5$	76 %	71 %	**70 %**	**72%**

in ages from 22-76. Nine different stress conditions were recorded: neutral, angry, loud, soft, slow, Lombard effect (pink noise presented binaurally at an 85 dB SPL level), fast, and speech produced under two levels of workload: low and high. Each class contains two tokens of 35 highly confusable aircraft communication words. For more information on the database the interested reader can refer to [15].

3.2. Benchmarks and Figure-of-Merit

As benchmarks, we have selected the baseline method used in recent emotion challenges. More specifically, models rely on the acoustic feature set extracted by the OpenSMILE toolkit and used in the INTERSPEECH 2010 Paralinguistic Challenge [21]. PCA is also applied to the utterance level features reducing dimensionality from 1582 to 200. As in this challenge and in our previous work, an SVM classifier is used for the benchmarks [22, 13], with a linear kernel. As figure-of-merit, classification accuracy on the unseen test set, for each of the two datasets, is reported. Three tasks are explored: (1) 2-class task (i.e., angry vs. neutral), (2) 4-class task (i.e., neutral, angry, soft, and fast), and (3) a 9-class task. Results are reported based on a 3-fold cross-validation.

4. EXPERIMENTAL RESULTS

Table 2 summarizes our findings for the 2-, 4-, and 9-class problems. As can be seen, the proposed method based on a CNN outperforms the other approaches for the most challenging task of 9-class classification. While it achieved 70 % accuracy, the benchmark provided 58 % (18 % lower accuracy). Notice that, the benchmark is also outperformed by our $\mathcal{E}_4$-DNN based method, which achieved 62 % accuracy for the same task, 4 % higher than the benchmark. Compared to all the other systems, the CNN based method showed to be the least affected when the number of classes were increased. The DNN-OpenSMILE system, for instance, provided the best results for the 2-class and 4-class tasks, but showed to be very sensitive as the number of classes increased. Its performance decayed 30% from the 2-class to 9-class problem while the proposed CNN system dropped only 10 %. On average (see last column), both methods presented equivalent accuracy 72%, with the $\mathcal{E}_4$-DNN based method being the second best system in average (68 % accuracy achieved).

The DNN-based systems increased the performance of all the features when compared to the SVM-based one. The gain for the OpenSMILE were substantial. We can observe in Table 2 an increase from 68 % to 83 %, for the 2-class problem, and from 57 % to 77 %, for the 4-class task, respectively 22 % and 25 % improvement. PCA was not applied prior to the DNN architecture. Although, it helps to decorrelate features and speed convergence it was less effective with the DNN.

Results in the Table show that our previous MSF features combined with a DNN model can be used for stress detection from speech, offering good performance especially in cases where the number of classes are considerably high. Moreover, as the performance of these features are dependent on the size of the analysis window used for the pooling procedure, from the results we can conclude that the proposed CNN method seems to be the optimal choice in lieu of the pooling scheme proposed in [13], as it improved performance across all the classification tasks.

5. CONCLUSION

The goal of this study has been two-fold. First, to explore the applicability of newly-proposed MSF and functionals-based pooling mechanisms for speech-stress detection. Second, to investigate the use of a the MSF features combined to a convolutional neural network to boost performance. We found that statistical pooling of MSF features can be useful for stress detection, but with performance constrained to utterance duration. By replacing the functionals-pooling step by a CNN, substantially higher accuracy is achieved. Two of the investigated methods outperformed the benchmark for the 9-class classification: the $\mathcal{E}_4$-DNN based system and specially the CNN-based one. Future work will explore fusion of the OpenSMILE features with that of MSFs to see if complementary can be found.

6. ACKNOWLEDGEMENT

The authors acknowledge NSERC, PROMPT, MITACS, THALES, CNPq and FRQNT for funding this work.

7. REFERENCES

[1] R. Li-Chern R.L Pan and J.K Li, "A noninvasive parametric evaluation of stress effects on global cardiovascular function," *Cardiovascular Engineering*, vol. 7, no. 2, pp. 74–80, 2007.

[2] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," *Computer methods and programs in biomedicine*, vol. 108, no. 3, pp. 1287–1301, 2012.

[3] O. Kofman and et al., "Enhanced performance on executive functions associated with examination stress: Evidence from task-switching and stroop paradigms," *Cognition & Emotion*, vol. 20, no. 5, pp. 577–595, 2006.

[4] O.V. Crowley and et al., "The interactive effect of change in perceived stress and trait anxiety on vagal recovery from cognitive challenge," *International Journal of Psychophysiology*, vol. 82, no. 3, pp. 225–232, 2011.

[5] T. Chandola, A. Heraclides, and M. Kumari, "Psychophysiological biomarkers of workplace stressors," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 1, pp. 51–57, 2010.

[6] S. Hafeez, "The impact of job stress on performance of employees: A study of social security hospital of district okara & sahiwal," *Journal of Neuropsychology & Stress Management*, vol. 3, pp. 4–12, 2018.

[7] H.G. Ásgeirsdóttir and et al., "The association between different traumatic life events and suicidality," *European journal of psychotraumatology*, vol. 9, no. 1, pp. 1510279, 2018.

[8] P. Rajasekaran, G. Doddington, and J. Picone, "Recognition of speech under stress and in noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.* IEEE, 1986, vol. 11, pp. 733–736.

[9] C.E. Williams and K.N. Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[10] C.K. Yogesh and et al., "Hybrid bbo_pso and higher order spectral features for emotion and stress recognition from natural speech," *Applied Soft Computing*, vol. 56, pp. 217–232, 2017.

[11] K. Wang and et al., "Speech emotion recognition using fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015.

[12] P. Tzirakis and et al., "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[13] A.R. Avila and et al., "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Transactions on Affective Computing*, 2018.

[14] A.R. Avila et al., "Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks," in *2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2017, pp. 360–365.

[15] J.H.L Hansen and S. E. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[16] S. Wu and et al., "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[17] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, pp. 8, 1993.

[18] S.D Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *Journal of Acoustical Society of America*, vol. 108, no. 3, pp. 1181–1196, 2000.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436, 2015.

[20] D.M. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[21] B. Schuller and et al., "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.

[22] Michel M.M. Valstar and et al., "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.