Interspeech 2022
18-22 September 2022, Incheon, Korea

State & Trait Measurement from Nonverbal Vocalizations: A Multi-Task Joint Learning Approach

Alice Baird [1], *Panagiotis Tzirakis* [1], *Jeffrey A. Brooks* [1], *Lauren Kim* [1], *Michael Opara* [1], *Christopher B. Gregory* [1], *Jacob Metrick* [1], *Garrett Boseck* [1], *Dacher Keltner* [1,2], *Alan S. Cowen* [1]

[1] Hume AI Inc., New York City, New York, USA
[2] University of California, Berkeley, California, USA
alice@hume.ai

Abstract

Humans infer a wide array of meanings from expressive nonverbal vocalizations, e. g., laughs, cries, and sighs. Thus far, computational research has primarily focused on the coarse classification of vocalizations such as laughs, but that approach overlooks significant variations in the meaning of distinct laughs (e. g., amusement, awkwardness, triumph) and the rich array of more nuanced vocalizations people form. Nonverbal vocalizations are shaped by the emotional state an individual chooses to convey, their wellbeing, and (as with the voice more broadly) their identity-related traits. In the present study, we utilize a large-scale dataset comprising more than 35 hours of densely labeled vocal bursts to model emotionally expressive states and demographic traits from nonverbal vocalizations. We compare a single-task and multi-task deep learning architecture to explore how models can leverage acoustic co-dependencies that may exist between the expression of 10 emotions by vocal bursts and the demographic traits of the speaker. Results show that nonverbal vocalizations can be reliably leveraged to predict emotional expression, age, and country of origin. In a multi-task setting, our experiments show that joint learning of emotional expression and demographic traits appears to yield robust results, primarily beneficial for the classification of a speaker's country of origin.

Index Terms: Affective computing, computational paralinguistics, nonverbal vocalizations, multi-task learning

1. Introduction

Nonverbal behavior is an essential component of human communication. What we say is often less important than the way we say it, including our emotional intonation, facial expression, and body language [1]. Computational studies of vocal emotional expression have focused on speech, but recent advances in the psychology literature suggest that nonverbal vocalizations such as laughs, cries, screams, sighs, and grunts may be an even more robust medium for the communication of emotion [2, 3]. From these brief bursts of sound, humans can distinguish a wide range of intended meanings as well as traits of the speaker [4].

Within the field of computational paralinguistics, few studies have explored nonverbal expressive vocalizations, with most focusing on the coarse classification of vocalizations such as laughs [5] or cries [6]. This limited focus is likely due to the limited availability of data capturing a wider variety of vocalizations and their more subtle meanings, along with the ease with which laughs and cries can be extracted from speech [7, 8].

However, the focus on laughs and cries overlooks significant variations in the meaning of distinct laughs and cries (e. g., laughs of amusement, embarrassment, or triumph) [2], let alone other common vocal bursts such as gasps. Understanding the wider array of emotional vocalizations has significant implications both for the study of social interactions – e. g., via back-channeling [9] and conversational rapport [10] – and for human-robot interaction [11], which has significant applications in interventions to help neurodiverse groups [12]. Furthermore, nonverbal vocalizations are often contagious [13], and understanding group expression has benefits for predicting community wellbeing.

Recent studies have established that nonverbal vocalizations convey a much wider range of emotional meanings than previously known, with at least 24 distinct dimensions of meaning including *Amusement*, *Distress* and *Awe* [2]. Only a small fraction of the nuanced, high-dimensional meanings that nonverbal vocalizations are now understood to convey are captured by the two traditional emotion models commonly applied in affective computing, including Ekman's basic six emotions [14] and the circumplex model of affect (valence and arousal) [15].

With this in mind, here we introduce a multi-task learning strategy to jointly learn the high-dimensional emotional meanings inferred from nonverbal vocalizations along with the demographic traits of the speaker, including age and country of origin. Multi-task learning has had been successfully applied to a range of feature-based speech-related tasks, including joint learning of multiple dimensions of emotion [16] and joint learning of demographic traits such as gender [17]. However, to the best of our knowledge, emotional expression and demographic traits have not yet been jointly learned using a multi-task approach, nor has any multi-task joint leaning approach been applied to nonverbal vocalizations.

In the present contribution, we conduct a series of deep learning-based single- and multi-task experiments, utilizing well-known audio feature-sets, along with the first-of-its-kind large-scale emotional nonverbal vocalization dataset HUME-VB. The core contributions of this work are two-fold; (1) For the first time, we computationally validate the efficacy of using nonverbal vocalization to measure and recognize emotionally expressed states and demographic traits. (2) We examine the benefits of jointly learning emotional expression and demographic traits from human vocalizations, and the implications for state- and trait-related processing architectures.

The paper is structured as follows: First, in Section 2 the dataset applied for these experiments is explained in detail, followed by a description of the data processing. We then detail our methodology in Section 3, including the features utilized and the architecture developed for single and multi-task learning. Proceeding this we discuss our results in Section 4 and conclude our findings, offering suggestions for future work in Section 5.

10.21437/Interspeech.2022-10927

Table 1: *An overview of* HUME-VB*, first presented in [18]. Including (No.) Samples, Duration (HH: MM: SS), and Speakers.*

	Train	Validation	Test	$\sum$
HH: MM: SS	12 :19 :06	12 :05 :45	12 :22 :12	36 :47 :04
No.	19 990	19 396	19 815	59 201
Speakers	571	568	563	1 702

2. The Hume Vocal Burst Dataset

For the experiments of the current contribution, we utilize the Hume Vocal Bursts (HUME-VB) dataset. This dataset was first presented as part of the 2022 ICML Expressive Vocalizations Workshop & Competition (EXVO) [18][1], and consists of 36:50:40 (HH :MM :SS), of audio data from 1,702 speakers (987 Female), aged from 20.5 to 39.5 years old. The subjects within the dataset come from 4 countries, the USA (21,544 samples), China (15,218), South Africa (14,321), and Venezuela (8,118), representing a broad range in culturally-derived characteristics. Furthermore, the data is collected in the subject's home via their own microphone, meaning that the data can be considered "in-the-wild" – consisting of uncontrolled conditions and noise profiles. Participants within HUME-VB were recruited via a range of crowdsourcing platforms (Amazon Mechanical Turk, Clickworker, Prolific, Microworkers, and RapidWorker). In each trial, participants heard a seed vocal burst and were instructed to use their computer microphone to record themselves mimicking the vocal burst such that their imitation would be perceived to convey similar emotions to the original recording. Participants completed 30 trials per survey and could complete multiple versions of the survey, up to 10 depending on the country. All participants provided informed consent and all aspects of the study were approved by Heartland IRB.

Each nonverbal vocalization in the dataset has been self-rated by the subjects themselves for their perceived emotional expression intensity (in a range of 1:100). Within this study, we focus on a selection of 10 emotional expressions; 1. Awe 2. Excitement 3. Amusement 4. Awkwardness 5. Fear 6. Horror 7. Distress 8. Triumph 9. Sadness 10. Surprise. The samples were selected from a broader database of vocal bursts that included 48 emotion ratings. The ten emotions selected were chosen based on their frequency within the dataset and to ensure the dataset included both broad distinctions (e. g., amusement vs. fear) and more nuanced distinctions (e. g., fear vs. horror), consistent with the idea that emotions occupy a high-dimensional continuous space.

In Figure 1, the distribution of emotional expressions across the entire training set is shown as a t-SNE representation. From Figure 1, we can see that in most cases, the expressions have a clear distinction within the embedding space, with clustering appearing to relate to conventional models of emotion. Of note, there are fewer samples for the 'Triumph' class, so we expected this class to be more challenging to model.

As a first step for processing the data, all of the audio files are normalized to -3 decibels, and converted to 16 kHz, 16 bit, mono. No other processing is applied to the files, although from a subjective overview of the data it may be of interest to explore the benefits of trimming the samples to remove areas of silence, amongst other strategies for speech enhancement from naturalistic environments.

[1]Access to the dataset used can be requested via Zenodo: https://doi.org/10.5281/zenodo.6308780

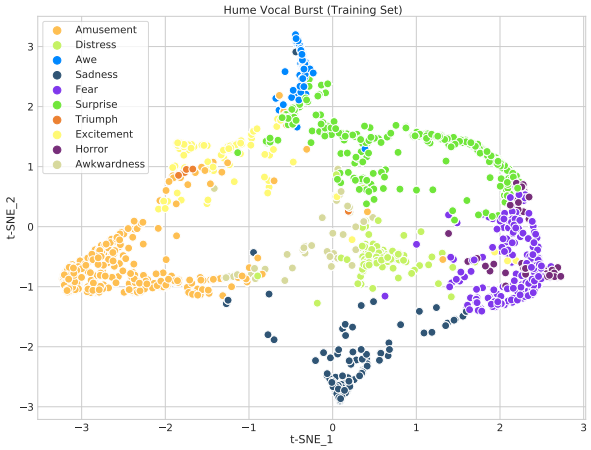

Figure 1: *t-SNE representation of the normalized emotional expression space for the* HUME-VB *training set.*

Proceeding this we partition the data into training, validation and test splits. In Table 1, the data quantity and distribution across the partition splits is provided. We have partitioned the data with consideration to speaker independence and therefore balanced the splits in a country-wise manner as well. As for the labels within the HUME-VB dataset, we normalize the emotion labels for maximum value per sample, to a range [0 :1] across each sample. The range of values for age are 20.5 to 39.5 years, and no further normalization is applied. There is no additionally processing made to the country-classes.

3. Methods

To explore the use of nonverbal vocalization for measuring a variety of state and trait attributes and the value of jointly modeling these attributes, we perform a series of experiments with both a single- and multi-task learning strategy. The architecture we apply follows the success of other speech-based multi-tasks strategies in affective computing [16], and we report model performance in terms of Concordance Correlation Coefficient (CCC) for the mean of the 10 emotional expressions, Mean Absolute Error (MAE) for age in years, and Unweighted Average Recall (UAR) for the country of origin classification task.

3.1. Audio Features

We extracted several well-established feature sets from the HUME-VB dataset for use within the single and multi-task learning pipelines. One feature vector is extracted per sample. Although it may be interesting to explore the benefit of smaller window sizes for feature extraction, the mean duration of the audio samples is reasonably short (2.23 seconds), so we leave this for future exploration. Standardization is also performed prior to training for each feature set. Utilizing the OPENSMILE toolkit, we extracted the 6 373-dimensional COMPARE set and the 88-dimensional EGEMAPS set. The COMPARE set has become a well-established feature set for paralinguistics tasks [19], and contains 6 373 static features which are the result of the calculation of functionals (statistics) from low-level descriptors (LLD) [20, 19]. The extended Geneva Minimalistic Acoustic Parameter Set (EGEMAPS) [21], is smaller in size and was designed for affective based computational paralinguistic tasks. Both sets continue to be valuable for several computational emotion-based speech studies [22, 23].

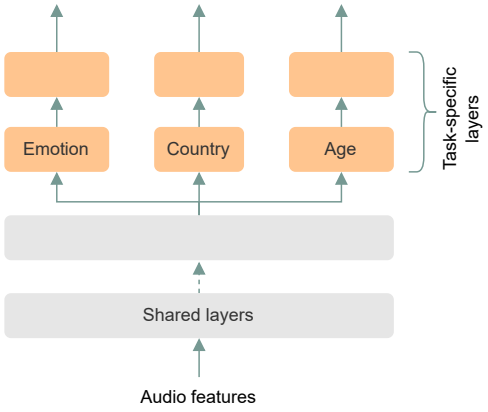

Figure 2: *Overview of the hard-parameter sharing multi-task learning architecture which is applied for these experiments.*

In addition to OPENSMILE, we also calculated a Bag-of-Audio-Words (BOAW) representation from the COMPARE LLD's, utilizing the OPENXBOW toolkit [24]. BOAW has been found to be extremely useful for speech-based emotion tasks [25], and essentially for each audio sample, a histogram of the acoustic LLD's, is created based on quantization set by a codebook. For our experiments, we calculated a codebook of size 2 000, based on the success in related works [26].

3.2. Network Architecture

To investigate the benefits of joint learning from states of emotional expression and demographic traits, we apply a multi-task learning (MTL) method, with a hard-parameter sharing deep learning architecture, in which the three tasks share the input layer, and are later fed to a task-specific output layers [27], see Figure 2 for a high-level schematic overview. We take inspiration for this from [16], in which the authors explore various MTL architectures, finding the hard-parameter sharing to be the most effective for their targets of emotional valence, arousal and dominance. Each layer in the network is a fully-connected one, where layer normalization is applied only to the shared layers. A leaky rectified linear unit (Leaky ReLU) is used as the activation function to all layers, as this was found to be the most effective during development likely due to its ability to combat the so-call 'dead neuron problem' [28]. To compare the efficacy of the multi-task strategy and explore how well nonverbal vocalizations can be applied to model these states and traits, we also run single-task experiments for each target. A similar network is utilized for the single-task models, excluding the redundant task-specific layers.

3.3. Model Training

The Adam optimization method is used for all experiments [29], and to calculate the loss for the multi-task model, we needed to consider an approach for combining the loss from each task's output layer. In this regard, several approaches within the literature have considered an 'uncertainty' weighting [30]. However, to avoid complexity in the results, for this contribution, we choose to apply only an equally weighted sum of each target's loss. For the emotion target, as this is a multi-label experiment, the Mean Square Error (MSE) loss is calculated for each class, then the average across the ten classes is calculated. For age, MSE is also used as the loss, and for the country of origin, we apply cross-entropy loss, which incorporates a softmax activation function as the last layer. After tuning the hyperparameters on the development set, we find that a learning rate (lr) of 10^{-3} and batch size (bs) of 8 is optimal. As we are explicitly comparing across features, targets and models, we consider that for this contribution reporting only a fixed lr and bs allows for greater interpretability of results. Furthermore, to avoid overfitting, we apply an early stopping strategy with a patience of 5 epochs, and a maximum number of epochs of 20.

Table 2: *Validation and Test results for the **single-task (STL) experiments**, targeting 10-class emotional expression, reporting mean CCC, 4-class country of origin, reporting UAR (chance 0.25 UAR), and age in years (range 20.5 to 39.5 years), reporting MAE. Reporting best score from 3 fixed seeds. All results are obtained with lr 10^{-3}, and bs 8. Emphasized results indicate the best on test for each task.*

STL		Val	±	Test	±
EGEMAPS	Emo CCC	0.3842	0.0035	0.3835	0.0025
	Cou UAR	0.4236	0.0055	0.4096	0.0065
	Age MAE	4.1798	0.0020	4.4943	0.0010
COMPARE	Emo CCC	0.4435	0.0041	0.4490	0.0046
	Cou UAR	0.5227	0.0031	**0.4996**	0.0039
	Age MAE	4.1127	0.0041	**4.3538**	0.0046
BOAW 2000	Emo CCC	0.4546	0.0039	**0.4612**	0.0046
	Cou UAR	0.4489	0.0038	0.4178	0.0016
	Age MAE	4.3521	0.0039	4.6091	0.0047

4. Discussion of Results

We have run a series of experiments to explore both the potential of using nonverbal vocalizations to predict emotional expression, age in years and country of origin, as well as the potential benefit of multi-task joint learning in this domain. Results for our single-task models are shown in Table 2, and multi-task experiments can be seen in Table 3. We will discuss the results for each first separately.

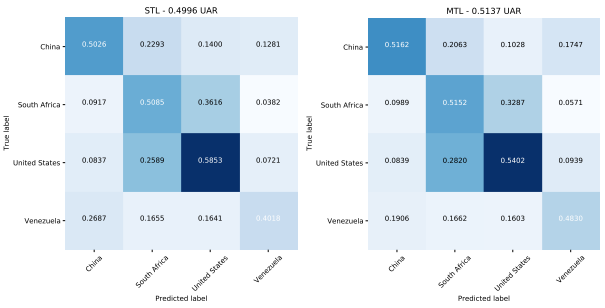

Figure 3: *Confusion matrix of results for the 4-class country of origin classification, with* COMPARE *features, in the STL (left) and MLT2 (right) paradigm.*

4.1. Single-task

In the single-task scenario (Table 2), we found that the features extracted from the nonverbal vocalizations can be used to model the three tasks of interest. Particularly for age, which has more of a benchmarkable history, we observe from the literature that

Table 3: *Validation and test results for the **multi-task (MTL) experiments**, targeting 10-class emotional expression, reporting mean CCC, 4-class country of origin, reporting UAR (chance-level 0.25 UAR), and age in years (range 20.5 to 39.5 years), reporting MAE. Results were obtained from the best score of 3 seeds, and a fixed lr 10^{-3}, with a bs 8, and no. of shared layers (S), and hidden units (H) is indicated. Emphasized results indicate the best on test for each task.*

		Val	±	Test	±
MLT 1: $S = 1$, $H = 64$					
ComParE	Emo CCC	0.4342	0.0020	**0.4405**	0.0014
	Cou UAR	0.5236	0.0134	0.4965	0.0097
	Age MAE	4.3468	0.0022	4.6695	0.0014
BoAW 2000	Emo CCC	0.4072	0.0014	0.4091	0.0005
	Cou UAR	0.4412	0.0078	0.4122	0.0060
	Age MAE	4.5268	0.0014	4.8600	0.0005
MLT 2: $S = 2$, $H = 128$					
ComParE	Emo CCC	0.4174	0.0011	0.4240	0.0028
	Cou UAR	0.5328	0.0113	**0.5137**	0.0099
	Age MAE	4.2706	0.0011	**4.5252**	0.0028
BoAW 2000	Emo CCC	0.4010	0.0048	0.4072	0.0073
	Cou UAR	0.4441	0.0041	0.4151	0.0035
	Age MAE	4.5467	0.0048	4.8657	0.0073
MLT 3: $S = 3$, $H = 256$					
ComParE	Emo CCC	0.3456	0.0095	0.3474	0.0094
	Cou UAR	0.5288	0.0138	0.5030	0.0116
	Age MAE	4.4135	0.0095	4.6129	0.0094
BoAW 2000	Emo CCC	0.3600	0.0310	0.3630	0.0306
	Cou UAR	0.4562	0.0131	0.4255	0.0111
	Age MAE	4.5201	0.0310	4.7888	0.0306

these experiments achieve a competitive MAE (ComParE features 4.35 MAE on test) as compared to similar speech-only based works [31]. We find the best results for country of origin by utilizing the ComParE features (0.4965 UAR), a score substantially above the chance level. However, as we can see from Figure 3, the class of Venezuela was more difficult to predict, and exploring strategies to compensate for this class imbalance, as well as the potential cultural bias which may be present from the feature sets, would be beneficial. As for the emotional expression scores, BoAW achieves the most robust score on the tests set (0.4612 CCC), with eGeMAPSnot performing as well for these tasks. Future work could explore the generation of feature sets specifically tailored to nonverbal vocalizations.

4.2. Multi-task

Given our findings from the STL experiments that ComParE and BoAW are the more optimal feature sets for these three tasks, we only report these sets for the MTL experiments. Across the MTL experiments, we were interested in exploring whether there are any advantages to incorporating multi-task joint learning to predict both emotional expression and demographic traits, as well as the effect of the number of shared layers across tasks on the models' ability to learn. As with the STL results, we observed that the models can generate robust predictions for all tasks (Table 3). Of particular interest, in this case, we see a slight improvement for MTL for predicting country of origin (ComParE 0.5137 UAR on test), although the improvement may be negligible given the standard deviation (0.0099) in this scenario. Nevertheless, as shown in Figure 3, we found that via the use of MTL, the minority class of Venezuela (0.4874 recall, vs. 0.3918 recall for STL) appears to be better predicted, with less confusion between Venezuela and the other classes.

In general, the use of MTL does not result in a substantial improvement in prediction accuracy over STL; however we note that 2 layers appear to offer more stable performance, with 3 layers beginning to degrade in performance, and show a much higher standard deviation across runs. Furthermore, from an efficiency perspective, it is worth noting that the MTL experiments required training three times fewer models than the STL experiments, with very little performance loss. Quantifying the efficiency benefits of MTL over STL in terms of training time and computing cost is out of scope for the present contribution but would be worth exploring in future work.

5. Conclusions

In the present contribution, we proposed using multi-task learning to explore the co-dependencies that may exist between emotional expression, country of origin, and age when learning from nonverbal emotional vocalizations. We sought to explore whether (1) models applied to human vocalizations could support accurate predictions of emotional expression and demographic traits and (2) multi-task joint learning of emotional expression and demographic traits would yield advantages over single-task learning. Our single-task results strongly support the use of nonverbal vocalizations to model emotional expression, country of origin, and age, with results for age being in a state-of-the-art range with consideration to speech only approaches. Multi-task joint learning of emotional expression and demographics does not appear to confer substantial improvements in prediction accuracy compared to STL. However, MTL does at least appear to better compensate for class imbalances, in this case, country of origin benefiting most from joint learning.

Based on the experiments herein there are several areas of interest which we hope will be explored in future work. From the the perspective of nonverbal vocalization modeling, we encourage further efforts to characterize the relationship between emotional expression and the acoustic features of vocalizations, including, but not limited to, the acoustic features which perform well in the present experiments. This would be particularly interesting to explore given the poor performance of the well-establish feature set used for speech-based tasks in affective computing, eGeMAPS. The poor performance of these features may indicate that the acoustic features relevant to verbal speech differ from those relevant to nonverbal bursts. Moreover, we believe the Hume-VB dataset could fruitfully be used to explore the possibility of vocalization enhancements. Additionally, as it pertains to MTL, and given that we do see some promise for joint learning of emotional expression and speaker demographics, we believe it could be valuable to explore in more detail how each dimension of emotional expression in nonverbal vocalizations interacts with acoustic properties related to each demographic trait, and to more precisely quantify potential benefits in efficiency (reduced overall training time and compute cost) when employing an MTL strategy.

6. References

[1] D. Phutela, "The importance of non-verbal communication," *IUP Journal of Soft Skills*, vol. 9, no. 4, p. 43, 2015.

[2] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 emotions conveyed by brief human vocalization." *American Psychologist*, vol. 74, no. 6, p. 698, 2019.

[3] J. S. Morris, S. K. Scott, and R. J. Dolan, "Saying it with feeling:

neural responses to emotional vocalizations," *Neuropsychologia*, vol. 37, no. 10, pp. 1155–1163, 1999.

[4] J. Bóna, "Non-verbal vocalizations in spontaneous speech: The effect of age," *The Phonetician*, vol. 115, pp. 23–35, 2018.

[5] K. P. Truong and D. A. v. Leeuwen, "Automatic detection of laughter," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[6] R. I. Tuduce, H. Cucu, and C. Burileanu, "Why is my baby crying? an in-depth analysis of paralinguistic features and classical machine learning algorithms for baby cry classification," in *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2018, pp. 1–4.

[7] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Springer, 2008, pp. 99–110.

[8] B. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, "The interspeech 2018 computational paralinguistics challenge: atypical and self-assessed affect, crying and heart beats," 2018.

[9] N. Moran, L. V. Hadley, M. Bader, and P. E. Keller, "Perception of 'back-channeling'nonverbal feedback in musical duo improvisation," *PLoS One*, vol. 10, no. 6, p. e0130070, 2015.

[10] G.-A. Levow and S. Duncan, "Contrasting cues to verbal and non-verbal backchannels in multi-lingual dyadic rapport," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[11] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *International Journal of Social Robotics*, vol. 11, no. 4, pp. 575–608, 2019.

[12] F. Marino, L. Ruta, D. Vagni, G. Tartarisco, A. Cerasa, and G. Pioggia, "Robot-assisted cognitive behavioral therapy for young children with autism spectrum disorders," *Encyclopedia of Autism Spectrum Disorders*, pp. 4004–4009, 2021.

[13] J. Trouvain and K. P. Truong, "Comparing non-verbal vocalisations in conversational speech corpora," in *Proceedings of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals*. Citeseer, 2012, pp. 36–39.

[14] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[15] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[16] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning." in *Interspeech*, vol. 2017, 2017, pp. 1103–1107.

[17] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition 'in the wild' using aggregated corpora and deep multi-task learning," *arXiv preprint arXiv:1708.03920*, 2017.

[18] A. Baird, P. Tzirakis, G. Gidel, M. Jiralerspong, E. B. Muller, K. Mathewson, B. Schuller, E. Cambria, D. Keltner, and A. Cowen, "The icml 2022 expressive vocalizations workshop and competition: Recognizing, generating, and personalizing vocal bursts," *arXiv preprint arXiv:2205.01780*, 2022.

[19] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. Interspeech*, Lyon, France, 2013, pp. 148–152.

[20] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. ACM Multimedia*, Barcelona, Spain, 2013, pp. 835–838.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "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, 2015.

[22] A. Baird, S. Amiriparian, and B. Schuller, "Can deep generative audio be emotional? towards an approach for personalised emotional audio generation," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.

[23] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The muse 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 2021, pp. 5–14.

[24] M. Schmitt and B. W. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[25] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proc. Interspeech*, San Francisco, CA, 2016, pp. 495–499.

[26] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen *et al.*, "The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates," *arXiv preprint arXiv:2102.13468*, 2021.

[27] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[28] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[31] M. Kaushik, T. T. Anh, E. S. Chng *et al.*, "End-to-end speaker age and height estimation using attention mechanism and triplet loss," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 1–8.