The NeurIPS 2023 Machine Learning for Audio Workshop: *Affective Audio Benchmarks and Novel Data*

Alice Baird*
Hume AI
New York, USA
alice@hume.ai

Rachel Manzelli*
Modulate AI
Massachusetts, USA
rachel@modulate.ai

Panagiotis Tzirakis
Hume AI
New York, USA

Chris Gagne
Hume AI
New York, USA

Haoqi Li
Hume AI
New York, USA

Sadie Allen
Boston University
Boston, USA

Sander Dieleman
DeepMind
London, UK

Brian Kulis
Boston University
Boston, USA

Shrikanth S. Narayanan
USC-SAIL
California, USA

Alan Cowen
Hume AI
New York, USA

Preliminary

Abstract

The NeurIPS 2023 Machine Learning for Audio Workshop brings together machine learning (ML) experts from various audio domains. There are several valuable audio-driven ML tasks, from speech emotion recognition to audio event detection, but the community is sparse compared to other ML areas, e.g., computer vision or natural language processing. A major limitation with audio is the available data; with audio being a time-dependent modality, high-quality data collection is time-consuming and costly, making it challenging for academic groups to apply their often state-of-the-art strategies to a larger, more generalizable dataset. In this short white paper, to encourage researchers with limited access to large-datasets, the organizers first outline several open-source datasets that are available to the community, for the duration of the workshop are releasing several propriety datasets. Namely, three vocal datasets, HUME-PROSODY, HUME-VOCALBURST, an acted emotional speech dataset MODULATE-SONATA, and an in-game streamer dataset MODULATE-STREAM. We outline the current baselines on these datasets but encourage researchers from across audio to utilise them outside of the initial baseline tasks.

1 Introduction

Working with audio data in machine learning presents unique challenges compared to fields like computer vision. Despite the importance of various key problems in the audio domain, such as text-to-speech, voice recognition, source separation, and synthesis, it has received considerably lower attention. However, there has been a recent renaissance in audio research, particularly in the field of synthesis, with the release of several influential papers in the last year [1–6]. The relative scarcity of prior research and this recent boom serves as the primary motivations behind organizing the 2023 NeurIPS Machine Learning for Audio (MLA) Workshop.

The workshop covers a vast scope of audio-related tasks, including but not limited to speech modeling, speech generation, music generation, denoising of speech and music, data augmentation, acoustic event classification, transcription, source separation, and even multimodal modelling involving audio.

*These authors contributed equally to this work.

There are prestigious competitions like the Detection and Classification of Acoustic Scenes and Events (DCASE) focusing on audio-driven machine learning [7], and numerous large-scale audio datasets with high-level labeling available in the literature, including Audioset [8], Urban80k [9], and Librispeech [10], along with unique concepts such as bird identification [11], and music genre detection [12]. Despite the availability of such datasets, there still exists a scarcity of openly accessible large-scale datasets particularly tailored for more specialized domains, such as human-computer interaction and human behavior analysis. The lack of specialized datasets presents a considerable hurdle for developing systems that understand more nuanced human characteristics.

To address this issue and to support authors submitting their work to MLA, the organizers have proactively provided four datasets to researchers participating in the workshop, namely, the HUME-PROSODY, HUME-VOCALBURST, MODULATE-STREAM and MODULATE-SONATA. These datasets offer a broad spectrum of human states, including both labeled and unlabeled data. Researchers are given unrestricted access to use these datasets for their submissions, providing them with the opportunity to tailor their strategies to new data situations and investigate innovative solutions. Moreover, in addition to utilizing the datasets for applied research, the workshop encourages researchers to test their approaches against established benchmarks on baseline tasks outlined in previous machine learning competitions [13, 14].

This paper aims to provide a comprehensive description of the four large-scale datasets made available in conjunction with this workshop (see Section 2). Additionally, for the datasets where applicable, we present benchmark results achieved thus far (see Section 3).

2 Workshop Audio Datasets

Researchers have harnessed an array of audio datasets to train models in various disciplines, including automatic speech recognition and speech diarization, among others. However, there exists a conspicuous scarcity of large-scale datasets specifically designed for speech emotion recognition and the study of people's responses to events. To bridge this gap, our workshop presents substantial datasets derived from individuals interacting with video games, making them ideally suited for investigating human responses to dynamic events. We further provide labeled datasets for emotion recognition, focusing on speech prosody and vocal bursts. These resources offer an invaluable benchmark for the evaluation and enhancement of unsupervised models, marking a significant advancement in the field of speech emotion recognition research. In what follows, we present four datasets: HUME-PROSODY (Sec. 2.1), HUME-VOCALBURST (Sec. 2.2), MODULATE-SONATA (Sec. 2.3), and MODULATE-STREAM (Sec. 2.4)

2.1 The Hume Expressive Prosodic Speech Dataset (HUME-PROSODY)

HUME-PROSODY represents a subset of a comprehensive dataset containing emotionally rated spoken utterances with diverse prosody. This subset comprises 41 hours, 48 minutes, and 55 seconds of audio data collected from 1,004 speakers, aged between 20 to 66 years old. The data was collected across three countries with distinct cultures: the United States, South Africa, and Venezuela. Notably, the data was recorded "in-the-wild", meaning it was captured in uncontrolled recording conditions using the speakers' own microphones.

The foundation of this dataset consists of more than 5,000 "seed" samples, which encompass various emotional expressions. These seed samples were gathered from openly available datasets such as MELD [15] and VENEC [16–18]. The seeds include a mix of 'same' sentences, such as over 500 instances of the phrase "Let me tell you something" [16], where the prosody plays a significant role in conveying meaning, and 'different' sentences, each with varying words and semantics, where the prosody's functional load is relatively lower.

Each audio sample in HUME-PROSODY is associated with intensity labels for ten different expressed emotions, ranging from 1 to 100. The complete Hume-Prosody dataset comprises 48 emotional expression dimensions, based on the semantic-space model for emotion [19]. However, for this particular subset, which was introduced in this year's Computational Paralinguistic Challenge (ComParE) [20], nine emotional classes were selected due to their more balanced distribution across the valence-arousal space. These classes include 'Anger,' 'Boredom,' 'Calmness,' 'Concentration,' 'Determination,' 'Excitement,' 'Interest,' 'Sadness,' and 'Tiredness.'

Preliminary

To create HUME-PROSODY, participants were recruited through various crowdsourcing platforms like Amazon Mechanical Turk, Clickworker, Prolific, Microworkers, and RapidWorker. They were instructed to mimic a seed vocal burst they heard and use their computer microphone to record themselves imitating the seed sentence with similar prosody to the original recording. Each participant completed 30 trials per survey, and they could complete multiple versions of the survey. The study received informed consent from all participants, and its design and procedure were approved by Heartland IRB.

The intensity ratings for each emotion were normalized to a scale ranging from 0 to 1. For baseline experiments, the audio files were normalized to -3 decibels and converted to 16 kHz, 16-bit, mono format (the raw unprocessed audio is also provided, captured at 48 kHz). No additional processing was applied to the files, making the data amenable to various tasks. Subsequently, the data was divided into training, validation, and test sets, ensuring speaker independence, see 1 for an overview.

Table 1: Summary of the HUME-PROSODY dataset, first presented at ComParE 2023 [20]. Including number of samples, speakers, and distribution of identified gender. Test split remains blind.

	Train	**Dev.**	**Test**
Sample no.	30,133	12,241	-
Speaker no.	600	202	-
Gender (f:m)	379:221	117:85	-

2.2 The Hume Expressive Vocal Bursts Dataset (HUME-VOCALBURST)

The HUME-VOCALBURST dataset is a vast collection of emotional non-linguistic vocalizations, also known as vocal bursts. It includes audio data totaling 36 hours, 47 minutes, and 04 seconds (HH:MM:SS) from 1,702 speakers aged between 20 and 39 years. The dataset was compiled in four countries with diverse cultures: China, South Africa, the U.S., and Venezuela. In these recordings, individuals imitated expressive seed samples, showcasing various emotions. The speakers' vocalizations were recorded in the comfort of their homes using their microphones.

Each vocal burst in the dataset has been rated by an average of 85.2 raters for the intensity of 10 different expressed emotions. These emotions are Amusement, Awe, Awkwardness, Distress, Excitement, Fear, Horror, Sadness, Surprise, and Triumph, each rated on a scale from 1 to 100. The intensity ratings for each emotion were scaled to a range of 0 to 1.

For the baseline experiments, the audio files were standardized by normalizing them to -3 decibels and then converted to a 16 kHz, 16-bit, mono format. Additionally, the dataset also provides participants with the original, unprocessed audio, which was captured at 48 kHz. To ensure fair evaluation, the data was divided into training, validation, and test sets, taking into account speaker independence and maintaining a balance across different emotion classes. (Refer to Table 2 for the details of the data partitioning.)

Table 2: Summary of the HUME-VOCALBURST data, first presented at ExVo 2022 [13]. Including number of samples, speakers, and distribution of identified gender. The age range for speakers is 20.5-39.5 years. For the purposes of the competition, the test split remains blind, speakers given for few-shot task.

	Train	**Dev.**	**Test**
Sample no.	19 990	19 396	19 815
Speaker no.	571	568	563
Gender (f:m)	305:266	324:244	—

2.3 The Modulate Acted Expression Speech Dataset (MODULATE-SONATA)

MODULATE-SONATA is an acted speech dataset, provided by Modulate. The dataset consists of 23 unique professional voice actors performing 15 total roles (more than 6 hours of audio data). Each

Preliminary

Table 3: Summary of the MODULATE-SONATA data. Including number of samples, speakers, and duration of audio data per partition split.

	Train	Dev.	Test
Sample no.	516	125	115
Speaker no.	16	4	3
Duration (HH:MM:SS)	04:11:03	00:59:04	01:01:54

`mp3` file contains the full audio recording of the actor speaking from a script of emotional sentences, in a particular *style* (role). The audio files contain the full script alongside a text file of the emotional sentence spoken at that specific timestamp.

There are 25 unique emotion classes in the dataset, including:

'*adoration*' , '*amusement*' , '*anger*' , '*awe*' , '*confusion*' , '*contempt*' , '*contentment*' , '*desire*' , '*disappointment*' , '*disgust*' , '*distress*' , '*elation*' , '*embarrassment*' , '*fear*' , '*hype*' , '*interest*' , '*pain*' , '*realization*' , '*relief*' , '*sadness*' , '*seductionecstasy*' , '*surprisenegative*' , '*surprisepositive*' , '*sympathy*' , '*triumph*'.

The 15 roles performed by the actors include anime voice archetypes (AVA), well-known actors, and fantasy characters. Specifically, these are:

'*AVABubblyAndSweet*' , '*AVANasallyAndMidpitched*' , '*AVAChild*' , '*BatmanImpression*' , '*EmmaWatsonImpression*' , '*JudyDenchImpression*' , '*KeanuImpression*' , '*KristenBellImpression*' , '*MatureAndSmokey*' , '*Nobility*' , '*Outlander*' , '*Pirate*' , '*ScarlettJohansonImpression*' , '*SmallCompanion*' , '*Spellcaster*' .

The script of emotional sentences read by the actors is provided in Appendix 5.1. For each emotional class there are between 2 to 6 sentences.

For the purposes of benchmarking, we segment the long-form audio files following the emotion labels. This results in 756 total audio samples, with an average of 29.52 seconds in length, and between 29-32 segments for each class. Proceeding this we have prepared a speaker-independent partition split (see Table 3 for further details), based on the individual segments of audio.

The audio was recorded using predominantly Neummann U87, SM7B, and Blue Yeti microphones, with some additional unknown condenser microphones used by actors who recorded their sessions remotely.

2.4 The Modulate Streamer Dataset (MODULATE-STREAM)

MODULATE-STREAM is an audio-only dataset of over 7 000 hours of publicly available gaming streams. The dataset is in total is 379GB in size, and contains 1 880 260 audio files saved in `opus` format for space efficiency.

Metadata is contained in a separate TSV file, and contains the following attributes for each audio file:

`[clip_name, player_id, session_id, clip_duration_msec, transcript, num_words, game_metadata]`

Each attribute is described as follows:

- `clip_name`: the audio filepath.
- `player_id`: an anonymized ID of the speaker. There are 940 unique players in this dataset.
- `session_id`: an anonymized ID of the stream. There are 2 839 unique streams in this dataset.
- `clip_duration_msec`: the clip duration in milliseconds.
- `transcript`: the approximate transcription of the audio contained within the clip, estimated by a `wav2vec`-based STT model.
- `num_words`: the number of words in the approximated transcription, with an average value of 20 for this dataset.

Preliminary

Table 4: Summary of the MODULATE-STREAM data. Including number of samples and speakers per partition split.

	Train	Dev.	Test
Sample no.	1 835 242	38 079	6 939
Speaker no.	658	141	141

- `game_metadata`: the game the speaker in the clip is playing in their stream (or if not a game, the general topic of the stream), if provided. There are 110 unique topics/games in this dataset.

3 Current Baselines, and Machine Learning Tasks

For the HUME-PROSODY and HUME-VOCALBURST datasets, various benchmarks are available, encompassing a variety of machine learning tasks. For MODULATE-SONATA we provide a simple baseline for speech emotion recognition to validate the efficacy of the datasets target domain. Due to the size of MODULATE-STREAM we do not provide any baseline results, however we have provided speaker-independent partition splits which anyone using the data are welcome to utilize.

In this section, we will provide an overview of the original baselines established by the workshop organizers for each dataset. Additionally, whenever applicable, we will showcase the best results achieved through contributions to previous competitions held where this data was provided. As HUME-PROSODY and HUME-VOCALBURST were provided as part of previously held competitions, the test sets for each of these will remain blind, and those wishing to evaluate for the tasks described herein should send predictions to `competitions@hume.ai`.

3.1 HUME-PROSODY Description of Tasks

The HUME-PROSODY have been assigned only one task, which is publicly available with detailed information provided by the ComParE challenge organizers [20]. You can find the baseline code for this task at `https://github.com/EIHW/ComParE2023`.

This specific task for the HUME-PROSODY is known as the *Emotion Share Sub-Challenge*, which involves a multi-label regression task. It requires participants to predict the proportion or 'share' of nine different emotions based on the ratings given by multiple raters for the 'seed' sample.

3.2 HUME-PROSODY Initial Baseline Results

For this baseline, embeddings were extracted from Wav2Vec2, fine-tuned to the MSP-Podcast dataset[21], to establish an initial baseline for the dataset. Next, a Support Vector Regressor was trained and evaluated, with the cost parameter C of the SVM optimized based on performance on the Dev set. After this optimization process, a final model was trained using the concatenated training and Dev sets for evaluation on the Test partition.

Table 5: Results for HUME-PROSODY. The official best results for Test are emphasized; Reporting Pearsons Correlation Coefficient across the mean of the available classes (ρ).

(ρ)	Dev.	Test	CI on Test
Wav2Vec2	.500	**.514**	.499 – .529
ComParE	.359	.365	.347 – .382
Late Fusion	.470	.476	.461 – .492

The competition is at this time still on-going and so we would advise the authors to check the literature relating to the ComParE 2022 challenge for any updated benchmarks.

Preliminary

3.3 HUME-VOCALBURST Description of Tasks

The HUME-VOCALBURST was first introduced during the ICML Expressive Vocal Burst (ExVo) Challenge [13], highlighting three key tasks with an emphasis on audio machine learning. Subsequently, it was presented at the Affective Vocal Burst (A-VB) Workshop at the 2022 ACII conference [14], where four additional tasks were defined with a focus on affective computing. In this section we will briefly describe each but please see the reference papers for further information.

ExVo Multi-Task Learning: Participants in this track will train multi-task models to predict 10 emotions, the speaker's age, and native country using vocal bursts. The baseline performance metric, is a combined metric (see equation 1) based on mean Concordance Correlation Coefficient (CCC) for emotions, Mean Absolute Error (MAE) for age, and Unweighted Average Recall (UAR) for native country, all of which will determine the final standings.

$$\mathrm{S_{MTL}} = \frac{3}{(1/\hat{\mathcal{C}} + 1/\hat{\mathcal{M}} + 1/\hat{\mathcal{U}})}. \tag{1}$$

ExVo Emotion Generation: This track requires teams to train generative models to produce vocal bursts for 10 distinct emotions. For evaluation in the competition combine metric was presented which included the quantitative methods of Fréchet Inception Distance (FID) [22]:

$$\mathrm{FID} = \|\mu - \mu^*\|^2 + Tr(C + C^* - 2(CC^*)^{1/2}), \tag{2}$$

As well as a qualitative approach based on human ratings (HEEP):

$$\mathrm{HEEP} = \sigma_{TH} / \sqrt{\sigma_T^2 \sigma_H^2}, \tag{3}$$

where T corresponds to the vectorized target matrix, a dummy matrix of size N (number of generated vocal bursts) by 10 (emotions), with ones for targeted emotions and zeros for non-targeted emotions, and H corresponds to the vectorized rating matrix, a matrix of size $N \times 10$ with entries corresponding to the average human intensity ratings of each generated vocal burst.

The two are then combined as S_{GEN}, and compute the mean between the inverted FID distance, and the HEEP score for each emotion (e); this is defined as

$$S_{GENe} = \frac{1/FID_e + HEEP_e}{2}. \tag{4}$$

Please note the human evaluation was provided by the organizer for the ExVo competition only.

ExVo Few-Shot Emotion Recognition: In this innovative track, teams will recognize emotional vocalizations using few-shot learning, emphasizing personalization by considering factors like pitch and frequency of the speaker's voice. Two labeled samples per speaker will be used for personalization.

A-VB High-Dimensional Emotion: Participants will predict the intensity of 10 specific emotions through a multi-output regression approach, evaluating their performance using the mean CCC across all emotions.

A-VB Two-Dimensional Emotion: Focused on predicting arousal and valence from the circumplex model of affect [23], participants will approach this task as a regression problem, reporting performance using the mean CCC across both dimensions.

A-VB Cross-Cultural Emotion: This unique track introduces a 10-dimensional emotion intensity regression task for four different countries, challenging participants to predict the intensity of 40 emotions. The evaluation will use the mean CCC across all 40 emotions.

A-VB Expressive Burst-Type: For this classification task, participants will aim to classify eight types of expressive vocal bursts (e.g., Laugh, Cry, Scream). Performance will be assessed using the Unweighted Average Recall (UAR), serving as an accuracy measure.

Preliminary

3.4 HUME-VOCALBURST Initial Baseline Results

For HUME-VOCALBURST there are several baselines set for the various tasks described. For both competitions the baseline code is provided at `https://github.com/HumeAI/competitions`.

For each competitions a variety of approaches were presented, including feature-driven [24], end-to-end [25] Long-short-term memory-based architectures, and Generative Adversarial Networks for generation.

Table 6: Results for the 2022 ICML ExVo Workshop [13]. Reporting best results for ExVo MultiTask, ExVo Generation, and ExVo FewShot Tasks.

(a) Development and baseline test scores for ExVo-MultiTask, reporting best for each as given in [13].

	Development				Test
	E-CCC	C-UAR	A-MAE	S_{MTL}	S_{MTL}
ComParE	**.416**	**.506**	4.22	**.349** $\pm$ .003	**.335** $\pm$ .002
eGeMAPS	.353	.423	**4.01**	.324 $\pm$ .005	.314 $\pm$ .005

(b) Fréchet inception distance (FID) for the ExVo-Generation task, for each emotions, no baseline was given for 'Triumph' (Tri.), as no samples were generated for this class.

	Amu.	Awe	Awk.	Dis.	Exc.	Fea.	Hor.	Sad.	Sur.	Tri.
FID	4.92	4.81	8.27	6.11	6.00	5.71	5.64	5.00	6.08	–
HEEP	0.49	0.46	0.04	0.32	0.08	0.04	0.27	-0.03	0.22	–
S_{GEN}	0.345	0.33	0.08	0.24	0.13	0.11	0.22	0.08	0.19	0.00

(c) Test results for the ExVo FewShot task. Reporting best score given in [13].

$\mathcal{H}$	Amu.	Awe	Awk.	Dis.	Exc.	Fea.	Hor.	Sad.	Sur.	Tri.	$\hat{\mathcal{C}}$
256	.554	.581	.282	.420	.311	.544	.490	.383	.561	.315	**.444** $\pm$ .006

3.5 HUME-VOCALBURST Latest Benchmark Results

As the competitions related to HUME-VOCALBURST we can provide the full list of results here for each, where applicable citing the relevant literature for the approach.

3.6 MODULATE-SONATA Description of Tasks

Given the high quality of the audio, and the scenario being acted, we suggest that the MODULATE-SONATA data be used for generation or as a benchmark for evaluation of speech emotion recognition tasks. With this in mind, we provide an initial speech emotion classification task utilizing all classes.

Table 7: Baseline scores for A-VB 2022 [13]. Reporting the mean Concordance Correlation Coefficient (CCC) for the three regression tasks and the Unweighted Average Recall (UAR) across the 8-classes for the vocal burst type task.

	CCC						UAR	
	High		*Two*		*Culture*		*Type*	
	Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test
ComParE [24]	.515	.521	.49¢	.499	.387	.380	.391	.384
END2YOU [25]	.564	**.569**	.499	**.508**	.436	**.440**	.417	**.417**

Preliminary

Table 8: Latest benchmarks for HUME-VOCALBURST based on the tasks presented in the 2022 ExVo Workshop. Best known result on the test set is emphasized. Reporting, S_{MTL} (see equation 1), S_{Gen} (see equation 4) and Concordance Correlation Coefficient (CCC), the task type respectively.

Team	**Task**	**Metric**	**Test**
Organisers [13]	MultiTask	S_{MTL}	.335
TeamAtmaja [26]	–	–	.378
IdiapTeam [27]	–	–	.379
EIHW-MM [28]	–	–	.394
0xAC [29]	–	–	.407
CMU_MLSP [30]	–	–	.412
NLPros [31]	–	–	**.435**
Organisers [13]	Generate	S_{GEN}	.090
Resemble [32]	–	–	.119
StyleMelMila [33]	–	–	**.408**
Organisers [13]	FewShot	CCC	.444
EIHW-MM [28]	–	–	.650
SaruLab-UTokyo [34]	–	–	**.739**

Table 9: Latest benchmarks for HUME-VOCALBURST based on the tasks presented in the 2022 A-VB Workshop. Best known result on the test set is emphasized. Report Concordance Correlation Coefficient (CCC), and Unweighted Average Recall (UAR) for the regression and classification tasks, respectively.

Team	**Task**	**Metric**	**Test**
Organizers [14]	High	CCC	.569
TeamEP-ITS [35]	–	–	.655
SclabCNU [36]	–	–	.668
HCAI [37]	–	–	.685
HCCL [38]	–	–	.724
Anonymous [39]	–	–	.730
EIHW [40]	–	–	**.736**
Organizers [14]	Two	CCC	.508
SclabCNU [36]	–	–	.620
TeamEP-ITS [35]	–	–	.629
HCCL [38]	–	–	.685
EIHW [40]	–	–	**.707**
Organizers [14]	Culture	CCC	.440
TeamEP-ITS [35]	–	–	.520
HCAI [37]	–	–	.526
SclabCNU [36]	–	–	.550
HCCL [38]	–	–	.602
EIHW [40]	–	–	**.620**
Organizers [14]	Type	UAR	.417
TeamEP-ITS [35]	–	–	.490
SclabCNU [36]	–	–	.497
EIHW [40]	–	–	.562
Team-AVB [41]	–	–	.519
HCAI [37]	–	–	**.586**

Preliminary

3.7 MODULATE-SONATA Initial Baseline Results

For MODULATE-SONATA we provide an initial baseline for the partitions provided. We extracted HuBERT and Wav2Vec2 embeddings utilizing the default parameters from each of the audio files, and took the mean across the time axis. For a classifier we utilize the `sklearn` Logistic Regression module, and optimize the value for C={0.001, 0.01, 0.1, 1} on the validation set, and retrain with the model with the validation and training set concatenated using on the best value for C, evaluated on the test set.

From the results in Table 10 we can see that Wav2Vec2 and HuBERT embeddings perform similarly for the emotion recognition task, and fusing these embeddings shows further improvement. In all cases, the strong performance across the data indicates the validity of this data for both speech emotion recognition and generation type tasks.

Table 10: MODULATE-SONATA validation and test sets performance, for 25 class, speech emotion recognition reporting Unweighted Average Recall (UAR), chance level 0.04.

Embedding	Validation	Test
Wav2Vec2	0.80	0.85
HuBERT	0.79	0.86
early fusion	0.82	0.90

4 Summary and Conclusions

In conclusion, for the 2023 NeurIPS Audio for Machine Learning Workshop we have made efforts to addresses the scarcity of specialized audio datasets by providing valuable resources, namely, HUME-PROSODY, HUME-VOCALBURST, MODULATE-SONATA, and MODULATE-STREAM. These datasets offer opportunities for researchers to explore diverse audio tasks, such as emotion recognition, vocal burst classification, speech generation, and unsupervised audio driven tasks. By establishing the outlined baselines and encouraging collaboration, the workshop aims fosters innovation in audio-driven machine learning, leading to the development of more robust models, advancing areas including human behavior understanding and human-computer-interaction amongst other.

References

[1] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation, 2022.

[2] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023.

[3] Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Text-to-music generation with long-context latent diffusion, 2023.

[4] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models, 2023.

[5] Kinyugo Maina. Msanii: High fidelity music synthesis on a shoestring budget, 2023.

[6] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

[7] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. Dcase 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

Preliminary

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[9] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

[10] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[11] Stefan Kahl, Amanda Navine, Tom Denton, Holger Klinck, Patrick Hart, Hervé Glotin, Hervé Goëau, Willem-Pier Vellinga, Robert Planqué, and Alexis Joly. Overview of birdclef 2022: Endangered bird species recognition in soundscape recordings. *Working Notes of CLEF*, 2022.

[12] Bob L Sturm. An analysis of the gtzan music genre dataset. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 7–12, 2012.

[13] Alice Baird, Panagiotis Tzirakis, Gauthier Gidel, Marco Jiralerspong, Eilif B Muller, Kory Mathewson, Björn Schuller, Erik Cambria, Dacher Keltner, and Alan Cowen. The icml 2022 expressive vocalizations workshop and competition: Recognizing, generating, and personalizing vocal bursts. *arXiv preprint arXiv:2205.01780*, 2022.

[14] Alice Baird, Panagiotis Tzirakis, Jeffrey A Brooks, Chris B Gregory, Björn Schuller, Anton Batliner, Dacher Keltner, and Alan Cowen. The acii 2022 affective vocal bursts workshop & competition. In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–5. IEEE, 2022.

[15] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *preprint arXiv:1810.02508*, 2018.

[16] Petri Laukka, Hillary Anger Elfenbein, Wanda Chui, Nutankumar S. Thingujam, Frederick K. Iraki, Thomas Rockstuhl, and Jean Althoff. Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. In *Proc. LREC 2010 Workshop on Corpora for Research on Emotion and Affect*, pages 53–57. LREC, Marrakesh, Morocco, 2010.

[17] Hillary Anger Elfenbein, Petri Laukka, Jean Althoff, Wanda Chui, Frederick K Iraki, Thomas Rockstuhl, and Nutankumar S. Thingujam. What do we hear in the voice? an open-ended judgment study of emotional speech prosody. *Personality and Social Psychology Bulletin*, 48(7):1087–1104, 2022.

[18] Petri Laukka, Hillary Anger Elfenbein, Nutankumar S Thingujam, Thomas Rockstuhl, Frederick K Iraki, Wanda Chui, and Jean Althoff. The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of personality and social psychology*, 111(5):686, 2016.

[19] Alan S Cowen and Dacher Keltner. Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, 25(2):124–136, 2021.

[20] Björn W Schuller, Anton Batliner, Shahin Amiriparian, Alexander Barnhill, Maurice Gerczuk, Andreas Triantafyllopoulos, Alice Baird, Panagiotis Tzirakis, Chris Gagne, Alan S Cowen, et al. The acm multimedia 2023 computational paralinguistics challenge: Emotion share & requests. *arXiv preprint arXiv:2304.14882*, 2023.

[21] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2019.

Preliminary

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 2017.

[23] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[24] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. the 18th ACM International Conference on Multimedia*, pages 1459–1462, 2010.

[25] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[26] Bagus Tris Atmaja, Zanjabila, and Akira Sasou. Jointly predicting emotion, age, and country using pre-trained acoustic embedding, 2022.

[27] Tilak Purohit, Imen Ben Mahmoud, Bogdan Vlasenko, and Mathew Magimai. Doss. Comparing supervised and self-supervised embedding for exvo multi-task learning track, 2022.

[28] Xin Jing, Meishu Song, Andreas Triantafyllopoulos, Zijiang Yang, and Björn W. Schuller. Redundancy reduction twins network: A training framework for multi-output emotion regression, 2022.

[29] Josh Belanich, Krishna Somandepalli, Brian Eoff, and Brendan Jou. Multitask vocal burst modeling with resnets and pre-trained paralinguistic conformers, 2022.

[30] Roshan Sharma, Tyler Vuong, Mark Lindsey, Hira Dhamyal, Rita Singh, and Bhiksha Raj. Self-supervision and learnable strfs for age, emotion, and country prediction, 2022.

[31] Atijit Anuchitanukul and Lucia Specia. Burst2vec: An adversarial multi-task approach for predicting emotion, age, and origin from vocal bursts, 2022.

[32] Chin-Cheng Hsu. Synthesizing personalized non-speech vocalization from discrete speech representations, 2022.

[33] Marco Jiralerspong and Gauthier Gidel. Generating diverse vocal bursts with stylegan2 and mel-spectrograms, 2022.

[34] Detai Xin, Shinnosuke Takamichi, and Hiroshi Saruwatari. Exploring the effectiveness of self-supervised learning and classifier chains in emotion recognition of nonverbal vocalizations, 2022.

[35] Bagus Tris Atmaja and Akira Sasou. Predicting affective vocal bursts with finetuned wav2vec 2.0, 2022.

[36] Dang-Khanh Nguyen, Sudarshan Pant, Ngoc-Huynh Ho, Guee-Sang Lee, Soo-Huyng Kim, and Hyung-Jeong Yang. Fine-tuning wav2vec for vocal-burst emotion recognition, 2022.

[37] Tobias Hallmen, Silvan Mertes, Dominik Schiller, and Elisabeth André. An efficient multitask learning architecture for affective vocal burst analysis, 2022.

[38] Jinchao Li, Xixin Wu, Kaitao Song, Dongsheng Li, Xunying Liu, and Helen Meng. A hierarchical regression chain framework for affective vocal burst recognition, 2023.

[39] Dang-Linh Trinh, Minh-Cong Vo, and Guee-Sang Lee. Self-relation attention and temporal awareness for emotion recognition via vocal burst, 2022.

[40] Vincent Karas, Andreas Triantafyllopoulos, Meishu Song, and Björn W. Schuller. Self-supervised attention networks and uncertainty loss weighting for multi-task emotion recognition on vocal bursts, 2022.

[41] Muhammad Shehram Shah Syed, Zafi Sherhan Syed, and Abbas Syed. Classification of vocal bursts for acii 2022 a-vb-type competition using convolutional neural networks and deep acoustic embeddings, 2022.

Preliminary

5 Appendix

5.1 MODULATE-SONATA Script

Amusement

"I'm the Juggernaut, bitch." "Want to know how I got these scars?...He turns to me and he says: Why. So. Serious. Let's put a smile on that face!" "In the face of overwhelming odds, I'm left with only one option. I'm gonna have to science the shit outta this."

Embarrassment

"I didn't mean to...just...just look away please!" "Are you kidding? I sounded like a total buffoon! I'm sure she'll never even want to talk to me again."

Sadness

"They cursed us. Murderer they called us. The cursed us, and drove us away. And we wept, Precious, we wept to be so alone. And we only wish to catch fish so juicy sweet. And we forgot the taste of bread...the sound of trees...the softness of the wind. We even forgot our own name. My precious." "Sometimes I wish I had never met you. Because then I could go to sleep at night not knowing there was someone like you out there." "We're going to be okay. You can rest now."

Elation

"I love the smell of napalm in the morning." "Carpe diem. Seize the day, boys. Make your lives extraordinary!" "Whew! That's the stuff!"

Triumph

"Are you not entertained! Are you not entertained! Is this not why you are here?!" "King Kong ain't got shit on me!" "We will not go quietly into the night! We will not vanish without a fight! We're going to live on! We're going to survive! Today we celebrate...our Independence Day!"

Contempt

"My name is Maximus Decimus Meridius, commander of the Armies of the North, General of the Felix Legions and loyal servant to the true emperor, Marcus Aurelius. Father to a murdered son. Husband to a murdered wife. And I will have my vengeance, in this life or the next." "You can torture us and bomb us and burn our districts to the ground. But do you see that? Fire is catching. And if we burn...you burn with us." "I'm surrounded by idiots."

Disgust

"Get your stinking paws off me, you damned dirty ape." "I can hardly forbear hurling things at him."

Disappointment

"Why is the rum gone? Why is the rum...always gone?" "Well look at you now; you just got your asses whipped, by a bunch of goddamn nerds. I'm sure your father is rolling his his grave." "I'm not angry with you. I'm just...disappointed."

Anger

"You talkin' to me? You talking to ME? ARE YOU TALKING TO ME? THEY CALL ME MR. PIG!" "Badges? We ain't got no badges! We don't need no badges! I don't have to show you any stinking badges!" "Enough is enough! I have had it with these motherfucking snakes on this motherfucking plane!"

Preliminary

Adoration

"When I drift off, I will dream about you. It's always you." "Because I still wake up every morning...and the first thing I want to do is see your face."

Sympathy

"Ye best start believin' in ghost stories, Miss Turner. Yer in one." "Oh yes, the past can hurt. But you can either run from it, or learn from it." "If I told you what happens...it won't happen."

Distress

"You don't understand! I coulda had class. I coulda been a contender. I could've been somebody, instead of a bum, which is what I am." "I don't know where I am. I don't know what's going no. I think I lost somebody but I..I can't remember..." "Gentlemen, you can't fight in here! This is the War Room!" "Westballz doesn't even try to win. He just styles on you as hard as possible, and he's so good at that, he just wins."

Pain

"I would have followed you...my brother...my captain...my king." "I don't care!" Harry yelled, snatching up a lunascope and throwing it into the fireplace. "I've had enough, I've seen enough, I want it to end, I don't care anymore!" "But anyways, I just - ouch! Oh, sonnoffabitch, right on the corner of the table, ooh, that one stings..."

Hype

"This is where we fight! This is where they die!" "Leeeeroy Jenkins!"

Seduction / Ecstasy

"Mm...this chocolate is rich...and velvety...and so very delicious..." "I bet you would enjoy that luscious dragonfruit very much, wouldn't you.." "Yes.. I want the cupcakes... I want all of the cupcakes." "Your appetite... is very impressive...." "The yogurt! It flows like water. It's soo good. I'll have what I'm having!"

Desire

"It is mine, I tell you. My own. My precious....yes, my precious..." "If I could just see her again...hear her again...touch her again...maybe I could die happy."

Contentment

"Mama always said life was like a box of chocolates. You never know what you're gonna get." "Elementary, my dear Watson." "Get busy living, or get busy dying." "No thanks, I'm perfectly happy sitting here with the sun on my face reading a good book for as long as the weather will let me."

Relief

"And...and I look at you, and I - I'm home." "Gentlemen, I wash my hands of this weirdness." "We're...we're alive? Oh, thank the merciful gods, we're alive."

Realization

"Houston, we have a problem." "You keep using that word. I do not think it means what you think it means." "Oh. Oh balls." "Oh. Em. Gee." "Oh my god. Did...did that just happen?" "Ahah! An act of true love can heal a frozen heart!"

Interest

"Hope. It is the only thing stronger than fear. A little hope is effective. A lot of hope is dangerous. A spark is fine, as long as it's contained." "What do you mean, you have a plan? Care to share?"

Preliminary

Confusion

“The plot thickens, as they say. Why, by the way? Is it a soup metaphor?” “What is this? A center for ants?”

Awe

“Toto, I’ve got a feeling we’re not in Kansas anymore.” “You’re gonna need a bigger boat.” “Yoo! Did he just walk up...slowly...and down smash?!”

Surprise (positive)

“Cinderella story. Outta nowhere. A former greenskeeper, now, about to become the Masters champion. It looks like a mirac - It’s in the hole! It’s in the hole! It’s in the hole!” “And the crowd’s cheering for a four stock...this could be it...THIS IS IT!”

Surprise (negative)

“You drilled a hole in the dentist?!” “What the - what the hell has been going on here? What, did you think I wouldn’t notice?”

Fear

“I’m not going near there! Sunnyside is a place of ruin and despair, ruled by an evil bear who smells of strawberries!” “You want me to fight? No way, man, he’s gonna kill me. He’s gonna kill me and my whole family. I’m out, you hear me, I’m out!”

Preliminary