ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095294

LARGE-SCALE NONVERBAL VOCALIZATION DETECTION USING TRANSFORMERS

Panagiotis Tzirakis [1], *Alice Baird* [1], *Jeffrey Brooks* [1,2], *Christopher Gagne*[1], *Lauren Kim* [1], *Michael Opara* [1], *Christopher Gregory* [1], *Jacob Metrick* [1], *Garrett Boseck* [1], *Vineet Tiruvadi*[1], *Björn Schuller*[3,4], *Dacher Keltner* [1,2], *Alan Cowen* [1,2]

[1] Hume AI Inc., New York City, New York, USA
[2] University of California, Berkeley, California, USA
[3] GLAM – Group on Language, Audio & Music, Imperial College London, UK
[4] EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

ABSTRACT

Detecting emotionally expressive nonverbal vocalizations is essential to developing technologies that can converse fluently with humans. The affective computing community has largely focused on understanding the intonation of emotional speech and language. However, advances in the study of vocal emotional behavior suggest that emotions may be more readily conveyed not by speech but by nonverbal vocalizations such as laughs, sighs, shrieks, and grunts – vocalizations that often occur in lieu of speech. The task of detecting such emotional vocalizations has been largely overlooked by researchers, likely due to the limited availability of data capturing a sufficiently wide variety of vocalizations. Most studies in the literature focus on detecting laughter or cries. In this paper, we present the first, to the best of our knowledge, nonverbal vocalization detection model trained to detect as many as 67 types of emotional vocalizations. For our purposes, we use the large-scale and in-the-wild HUME-VB dataset that provides more than 156 h of data. We thoroughly investigate the use of pre-trained audio transformer models, such as Wav2Vec2 and Whisper, and provide useful insights for the task at hand using different types of noise signals.

***Index Terms*—** Nonverbal vocalization, transformers, vocal burst detection

1. INTRODUCTION

Recognizing vocal emotional behavior is critical to technologies intending to converse fluently with humans and anticipate our needs, such as digital assistants and therapeutics. The human voice can convey emotion nonverbally in two ways [1]: through speech prosody (the tune, rhythm, and timbre of speech) and nonverbal vocalizations [2, 3]. Whereas prosody interacts with the lexical components of speech to convey emotions [2], nonverbal vocalizations (or vocal bursts), such as laughs, sighs, shrieks, grunts, moans, and roars, occur in the absence of speech [4].

Within the audio domain, the affective computing field has largely focused on speech prosody [5, 6, 7, 8, 9], but recent evidence suggests that emotions are even more readily conveyed by vocal bursts [3]. The task of understanding the broader array of emotional vocalizations has been largely overlooked by researchers and developers, likely due to the limited availability of data capturing a sufficiently wide variety of vocalizations to train and test machine learning models. Vocal bursts are simultaneously difficult and critical to capture because they represent a small percentage of human vocalization compared to speech, yet when they do occur, they are likely to set the tone of the interaction [10]. Thus, studies that model vocal bursts by first detecting them in speech using datasets designed for speech emotional expression recognition purposes have had limited success in developing accurate models for vocal bursts. These studies have generally focused on the coarse classification of certain types of vocal bursts, such as laughter [11, 12], but these approaches overlook significant variations in the meaning of distinct laughs (e. g., amusement, embarrassment, triumph) [3], let alone other common vocal bursts such as sighs, gasps, huhs, and ohhs.

Our focus in this study is on detecting emotionally expressed nonverbal vocalizations in an end-to-end manner under unconstrained environmental conditions, such as background noises, and invariance to the voice characteristics of speakers. For our purposes, we use the Hume Vocal Bursts dataset (HUME-VB), a large-scale and in-the-wild dataset that comprises of more than 194 hours ($280,000$ samples) from more than $3,500$ individuals. A subset of this dataset was used in the recent ICML ExVo [13] and ACII A-VB [14, 15] competitions. The breadth of vocal bursts represented in HUME-VB is unprecedented, providing ratings for 48 dimensions of emotional expression (e. g., Love, Amusement, Embarrassment, and Anger), along with 67 classes that describe the type of vocal bursts (e. g., Cry, Gasp, Giggle, Ooph, Ouch, and Oww). The recordings were collected within speakers' homes via their own microphones under uncontrolled environmental conditions and in realistic

settings. Speakers are from 5 countries spanning 4 native languages – the USA (English), China (Mandarin), Venezuela (Spanish), South Africa (English), and India (Hindu). To our knowledge, HUME-VB is the largest nonverbal vocalization dataset focusing on emotional expressions in the literature.

In this paper, we investigate the use of state-of-the-art pretrained transformer-based architectures to detect vocal bursts. Such models have been shown to provide great results in predicting the emotional response of vocal bursts [16]. For our purposes, we experiment with the recently proposed Whisper model [17] and Wav2Vec2 [18]. Both models have been trained on hundreds of hours of audio data, making them excellent models for the task at hand, given their high cross-domain generalisability. To the best of our knowledge, this is the first large-scale study focusing on nonverbal vocalization detection of emotional expression and the first to use the Whisper model in this setting.

2. THE HUME VOCAL BURST DATASET

The Hume Vocal Burst (HUME-VB) dataset consist of 194 :26 :35 (HH :MM :SS) of total audio data from $4,080$ individuals aged from 18 to 92 years old, and from five countries: United States, Venezuela, India, China, and South Africa, making the data culturally diverse. The audio samples were recorded in uncontrolled environmental conditions, and in particular, at the subjects' homes via their own microphones, and as such provides "in-the-wild" recording characteristics with different room reverberation (e. g., living room, bedroom), microphones (e. g., laptop, tablet, desktop), and background noises (e. g., street noise, kids playing).

Participants were recruited via a range of crowdsourcing platforms such as Amazon Mechanical Turk, Clickworker, Prolific, Microworkers, and RapidWorker. Participants heard an audio recording of a "seed" vocal burst and were instructed to record themselves mimicking the vocal burst such that their imitation by conveying similar emotions to the "seed" recording. Participants completed 30 trials per survey and could complete multiple versions of the survey. All participants provided informed consent and all aspects of the study design and procedure were approved by the Heartland IRB.

The dataset provides labels for 67 dimensions for describing the type of the vocal bursts (Table 1). Each participant selected the description that applies best to their own recording (self-report rating) and to five different recordings (perceptual rating). In total, $270\,576$ ratings are provided in the dataset, with each vocal burst to have been annotated by 1.34 raters on average.

We split our data into training, validation, and test sets in a subject-independent manner and based on the recording type (e. g., mimic or seed). In more detail, our training and validation sets contain all the mimic samples, and the test set all the seed samples. We perform this split as the seed samples are "clean" recordings without background noise, and mixing them with noise samples will provide more "natural" sounds. Table 2 shows the statistics of our splits and the dataset.

"Cackle", "Cheer", "Chuckle" "Cry", "Gasp", "Giggle", "Groan", "Growl", "Grunt", "Hiss", "Hoot", "Howl", "Laugh", "Moan", "Pant", "Roar", "Scream", "Screech", "Shout", "Shriek", "Sigh", "Snicker", "Snort", "Sob", "Squeal", "Wail", "Wheep", "Whimper", "Yawn", "Yelp", "Ah", "Aha", "Ahh", "Argh", "Aww", "Eek", "Eww", "Grr", "Ha", "Hah", "Haha", "Hehe", "Hmm", "Huh", "Hurray", "Mhm", "Mmm", "Oh", "Ohh", "Ooh", "Ooph", "Ouch", "Oww", "Pff", "Phew", "Tsk", "Ugh", "Uh", "Uh-huh", "Umm", "Whee", "Whew", "Woah", "Wow", "Yay", "Yippee", and "Yuck"

Table 1. HUME-VB Description Labels

	Train	Validation	Test	$\sum$
HH: MM: SS	130:26:41	26:25:23	1:11:51	157:53:56
# Samples	240 832	45 857	2 751	289 445
Speakers	2 951	593	—	—
F:M:O	1488:1342:121	325:250:18	—	—
USA	1227	228	—	—
India	329	69	—	—
China	320	62	—	—
South Africa	902	204	—	—
Venezuela	173	30	—	—

Table 2. An overview of the HUME-VB data. Including (No.) Samples, Duration (HH: MM: SS), Speakers, (M)ale:(F)emale:(O)ther, and per native-country. For the purposes of the competition, the test set is blinded.

3. AUDIO TRANSFORMER MODELS

Audio transformer models have been shown to provide state-of-the-art results in Automatic Speech Recognition (ASR) [18, 17]. Although these models were trained for ASR, the audio community has leveraged them in a number of different tasks including but not limited to speech emotion recognition [19] and vocal burst emotion recognition [20, 21]. This study focuses on the two most prominent transformer models, the Wav2Vec2 [18] and the Whisper model [17].

3.1. Wav2Vec2

Wav2Vec2 [18] is a transformer-based architecture trained in a self-supervised manner using speech signals. The input to the model is the raw waveform which is first processed by a multi-layer convolutional feature encoder. The encoded signal is then fed to a transformer model that builds a representation for the entire sequence with a 20 ms frame rate. This representation is then discretized and passed in parallel through

Model	# Train data (h)	# Params. (M)
Whisper-Tiny	$680k$	7.6
Whisper-Base	$680k$	19.8
Whisper-Small	$680k$	84
Wav2Vec2-Base	960	95

Table 3. Transformer models. Showing the number of hours the models were trained and their corresponding number of parameters.

a context network, and these discretized tokens are used as self-supervised labels during training.

Different model types are provided based on the size of the model (e. g., Base and Large) and the size of the dataset on which the model was trained. For our purposes and following our initial plan to have a small-size model, we experiment with the Wav2Vec2-Base model. We average the output frames across the sequence dimension in order to have a single feature vector for each input sequence.

3.2. Whisper

Whisper [17] is a sequence-to-sequence transformer model that was trained in a weakly supervised manner in a multitask setting. Tasks used for training the model include transcription, translation, and speech detection. The input to the model is a 30 s window which is first transformed into a log-Mel spectrogram with 80 channels and a 25 ms window. The output representation has the same frame rate as the input.

Four model types are provided: tiny, base, small, and large. To use these models for our task, we exploit the output of the encoder network and change the input window of the model to be of any length. We investigate the tiny, base, and small models. Similar to Wav2Vec2, we average all output frames across time to get a single representation for the input sequence.

4. TRAINING PROCESS

Noise Datasets. One of our goals is to train a model that performs well on several different environmental conditions. As such, it is important to train it with different noise types such that the model would be able to differentiate between vocal bursts and other audio sounds. To this end, we used part of the AudioSet [22], a large scale dataset with more than 500 classes. For our purposes, we downloaded 300 k (832 h) samples that span all of the classes. Although the transformer models were pre-trained using speech signals, we found that the out-of-the-box models provided high rate of false positives with speech signals as input. Therefore, we populate our noise dataset with speech signals from the LibriSpeech-train-clean-360, AudioSet-train, and Hume-Prosody-train dataset, which contains speech signals with different emotional expressions. Our validation set is comprised of the Libry-Party, LibriSpeech-dev-clean, RAVDESS, AudioSet-valid, and Hume-Prosody-valid dataset. For the Hume-Prosody dataset different subjects are used for training and validation. For the AudioSet samples, the noise types can be the same between training and validation, but with different samples. We did this because we wanted our model to "see" as many noise types as possible. Table 4 provides a summary of the noise datasets used in training and validation sets. Our test noises comprises of the datasets: LibriSpeech-test, Device and produced speech (DAPS), Chime, and ESC-50, with the total duration to be 89 h 34 min.

Dataset	Train	Validation	Total
AudioSet	$250k$ (694)	$50k$ (138)	$300k$ (832)
Hume-Prosody	$100k$ (80)	$20k$ (17)	$120k$ (97)
LibriSpeech	$104k$ (363)	8 703 (25)	$111k$ (387)
LibriParty	–	350 (28)	350 (28)
RAVDESS	–	2 452 (3)	2 452 (3)
Total	$454k$ (1 137)	$81.1k$ (211)	$533.2k$ (1347)

Table 4. Noise datasets. Showing the number of samples and in parenthesis the corresponding hours for the training and validation sets.

Sampling Strategy. It is important how we create our batches as we want our model to be trained with a similar amount of vocal bursts and noise samples. To this end, we uniformly sample vocal bursts and noise signals. In addition, our vocal burst signals are further uniformly sampled based on their description. As most of our noise samples are speech, we uniformly sample speech and noise types signals.

Augmentation. To further improve the detection quality of our model, we augment and mix the selected signals with different noise types such that the model is able to detect a vocal burst in different signal-to-noise (SNR) levels. As our augmentation methods, we used adding random Gaussian noise with the amplitude in the range $[.001, .0015]$, stretching the input signal with a rate in the range $[.8, 1.25]$, shifting the pitch in the range $[-4, 4]$ seminotes, and masking the time in the range $[.05, .25]$. A sample is augmented with at least two of these methods, with maximum four. Finally, after augmenting the input signal, we mix it with a random noise type using SNR levels in the range $[-10, 30]$.

Input. The input to the models during training is fixed to 2.5 s with all audio samples to be re-sampled to 16 kHz. In the case where a sample is longer, we trim it in a random location, and if a sample is shorter, we randomly pad it (left or right). The choice of the input length is based on (1) the average duration of the vocal burst and (2) it acts as an augmentation approach

as it introduces some randomness in the input samples.

5. EXPERIMENTS

5.1. Experimental Setup

We fine-tuned the transformer models by using a single neuron and a sigmoid activation function on top of the transformer embedding. For our purposes, we utilize the Adam optimization algorithm, and a learning rate of 10^{-4} for the output layer and 10^{-5} for the transfomer models with a mini-batch of 16 samples throughout all experiments. The weights of the output layer have been initialized with Kaiming uniform initialization, and a gradient norm clipping of 1.0 is used. Finally, the models output a single value per input segment, indicating if it contains a vocal burst or not. As our training loss, we used binary cross entropy.

5.2. Results

We first compare the different transformer models using all datasets in the test set. Of particular interest to us is the model's performance with speech samples, as these are more realistic conditions. As such, we split the noise datasets into those that contain speech and those that do not. The evaluation is performed on a sample level where, if a vocal burst is detected in an audio sample that does not contain one, then the whole sample is labeled as false positive. The samples that contain vocal bursts are the seed samples of our dataset (Table 2). Table 5 shows the results using the F1 score and Unweighted Average Recall (UAR) as our metrics.

We see that the Whisper-Small model performs best, with a small performance gap with the other models. Surprisingly even the Whisper-tiny model with only 7.6 M parameters performs excellent for the task at hand, with comparable performance with Wav2Vec2 and Whisper-Small. We also observe that Wav2Vec2-base performs worse than Whisper-Small and is similar to Whisper-Base, even though the later models have fewer parameters. Especially, Whisper-base has only a quarter (20 M) parameters compared to Wav2Vec2-base. This was expected, as the Whisper models were trained on $680k$ h of data compared to Wav2Vec2, which was trained on 960 h.

Diving deeper into the non-speech test dataset, we searched for the noise types with the highest number of false positives. The top three noise types were coughing, animal vocalization, and wind. Coughing is a type of nonverbal vocalization, but it does not provide emotional expression. Interestingly, it is not present in our dataset. Animal vocalization can be easily confused with human vocalization, even for the human ear. Finally, we believe that wind is confused with the "Pff" vocalization.

Of particular interest to us is the models' performance on conversations. To this end, we use the LibriSpeech-test data to create conversations by combining multiple (max 2) input files with a single (randomly selected) vocal burst. The duration of the dataset is in the range $[.5, 15]$ sec. A vocal burst can be anywhere in the audio samples, meaning it can also overlap with the speech signal. The vocal bursts and the speech signals are mixed with different SNR levels ({-5, 0, 5, 10}) in order to test the resilience of the models on different conditions. In total, a dataset with 2 h and 12 min is created. To run our models, we first segment the input signal to 2.5 sec (non-overlapping) windows. Table 5.2 shows the results. We consider a true positive sample ones where it is detected with the start and end boundaries to be 50 ms close to the ground truth [12]. From the results, we observe that Whisper-Small outperforms on average Wav2Vec2-base. Furthermore, we observe a low performance for the Whisper-Tiny/Base models when the SNR is 0 or lower. On the other hand, the larger models are more resilient to low SNR values.

Model	**F1** [%] ↑	**UAR** [%] ↑
Whisper-Tiny	91.8	92.6
Whisper-Base	93.4	93.7
Whisper-Small	**96.2**	**94.7**
Wav2Vec2-Base	94.6	93.9

Table 5. Results with respect to F1 (%) and UAR (%) scores on the test set.

Model	**SNR Level (dB)**			
	-5	0	5	10
Whisper-Tiny	52.2	76.8	82.3	87.9
Whisper-Base	58.8	78.3	86.8	89.5
Whisper-Small	**78.4**	90.7	**91.6**	**91.8**
Wav2Vec2-Base	76.9	**91.1**	89.3	89.7

Table 6. Results with respect to F1 [%]↑ score on the created mixed dataset on different SNR levels (-5, 0, 5, 10).

6. CONCLUSIONS

In this paper, we introduced a large-scale study for detecting nonverbal emotional vocalizations. For our purposes, we used the HUME-VB dataset that provides more than 150 h of data, and we investigated the use of pretrained audio transformers. We showed that the Whisper model outperforms Wav2Vec2 for the task at hand, even with a smaller number of parameters. In addition, we showed that the three noise types that provide a high number of false positives are coughing, wind, and animal vocalization. Finally, we investigated the models' performance when vocal bursts co-occur with speech signals using different SNR levels. Interestingly, we found that the larger models, namely, Whisper-Small and Wav2Vec2 are resilient to low SNR values, whereas the smaller models are not. In the future, we intend to explore detecting nonverbal vocalizations that occur simultaneously (e. g. people shouting).

7. REFERENCES

[1] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*, John Wiley & Sons, 2013.

[2] D. Keltner, J. Tracy, D. A. Sauter, D. C. Cordaro, and G. McNeil, "Expression of emotion," *Handbook of emotions*, vol. 4, pp. 236–249, 2016.

[3] A. Cowen, D. Sauter, J. L Tracy, and D. Keltner, "Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 69–90, 2019.

[4] R. Mitchell and Elliott D. R., "Attitudinal prosody: what we know and directions for future study," *Neuroscience and biobehavioral reviews*, vol. 37, pp. 471–479, 2013.

[5] P. Tzirakis, J. Zhang, and B.W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5200–5204.

[6] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[7] 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.

[8] 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.

[9] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, "Speech emotion recognition considering nonverbal vocalization in affective conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1675–1686, 2021.

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

[11] R. B Kantharaju, F. Ringeval, and L. Besacier, "Automatic recognition of affective laughter in spontaneous dyadic interactions from audiovisual signals," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 220–228.

[12] S. Condron, G. Clarke, A. Klementiev, D. Morse-Kopp, J. Parry, and D. Palaz, "Non-Verbal Vocalisation and Laughter Detection Using Sequence-to-Sequence Models and Multi-Label Training," in *Proc. Interspeech 2021*, 2021, pp. 2506–2510.

[13] 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.

[14] A. Baird, P. Tzirakis, J. A. Brooks, C. B Gregory, B. Schuller, A. Batliner, D. Keltner, and A. Cowen, "The acii 2022 affective vocal bursts workshop & competition: Understanding a critically understudied modality of emotional expression," *arXiv preprint arXiv:2207.03572*, 2022.

[15] B. Tris A. and A. Sasou, "Predicting affective vocal bursts with finetuned wav2vec 2.0," 2022, arXiv.

[16] D. Xin, S. Takamichi, and H. Saruwatari, "Exploring the effectiveness of self-supervised learning and classifier chains in emotion recognition of nonverbal vocalizations," *arXiv preprint arXiv:2206.10695*, 2022.

[17] A. Radford, J W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022, arXiv.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[19] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech 2021*, pp. 3400–3404, 2021.

[20] A. Anuchitanukul and L. Specia, "Burst2vec: An adversarial multi-task approach for predicting emotion, age, and origin from vocal bursts," *arXiv preprint arXiv:2206.12469*, 2022.

[21] M. I. B. Vlasenko B. Magimai-Doss M. Purohit, T., "Comparing supervised and self-supervised embedding for exvo multi-task learning track," *arXiv preprint arXiv:2206.11968*, 2022.

[22] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.