Transformer-based Non-Verbal Emotion Recognition: Exploring Model Portability across Speakers' Genders

Transformer-based Non-Verbal Emotion Recognition: Exploring Model Portability across Speakers' Genders

Lorenzo Vaiani*
lorenzo.vaiani@polito.it
Politecnico di Torino
Turin, Italy

Alkis Koudounas*
alkis.koudounas@polito.it
Politecnico di Torino
Turin, Italy

Moreno La Quatra*
moreno.laquatra@polito.it
Politecnico di Torino
Turin, Italy

Luca Cagliero
luca.cagliero@polito.it
Politecnico di Torino
Turin, Italy

Paolo Garza
paolo.garza@polito.it
Politecnico di Torino
Turin, Italy

Elena Baralis
elena.baralis@polito.it
Politecnico di Torino
Turin, Italy

ABSTRACT

Recognizing emotions in non-verbal audio tracks requires a deep understanding of their underlying features. Traditional classifiers relying on excitation, prosodic, and vocal traction features are not always capable of effectively generalizing across speakers' genders. In the ComParE 2022 vocalisation sub-challenge we explore the use of a Transformer architecture trained on contrastive audio examples. We leverage augmented data to learn robust non-verbal emotion classifiers. We also investigate the impact of different audio transformations, including neural voice conversion, on the classifier capability to generalize across speakers' genders. The empirical findings indicate that neural voice conversion is beneficial in the pretraining phase, yielding an improved model generality, whereas is harmful at the finetuning stage as hinders model specialization for the task of non-verbal emotion recognition.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; **Supervised learning by classification**; **Learning latent representations**.

KEYWORDS

Non-verbal emotion recognition, Audio classification, Contrastive learning, Data augmentation

1 INTRODUCTION

Emotion Recognition is a sentiment analysis subtask, which has recently received much attention by the Natural Language Understanding community [1]. For example, extracting and classifying emotions has shown to be very important in the study Human Computer Interactions [19] and in the development of in-vehicle monitoring systems [34] and psychological diagnosis tools [14].

The Vocalisations sub-challenge of the *Com*putational *Par*alinguistic Challeng*E* (*ComParE*) [35] entails recognizing emotions (i.e., achievement, anger, fear, pain, pleasure, surprise) from non-verbal vocal expressions. It addressed two main challenges:

(1) Tackling the emotion recognition task in absence of verbal speech content, which prioritizes the analysis of paralinguistic information.

(2) Overcoming the limitation of traditional paralinguistic models in coping with speakers with different characteristics. Specifically, ComParE focuses on generalizing the emotion classification task across speakers' genders. The purpose is to learn an audio classification model, leveraging only female voices, that is able to generalize on male vocalisations as well.

We address the ComParE task using the established transformer architecture [42], which entails pretraining a general-purpose model on a large dataset and then finetuning it for the non-verbal emotion recognition task.

To make emotion classifiers more portable to different speakers' genders we leverage data augmentation strategies and contrastive learning techniques to generate effective audio representation. Specifically, both pretraining and finetuning steps also consider altered versions of the original audio recordings. The applied transformations include both classical acoustic signal alterations (e.g., pitch shifting [29]) and more advanced neural voice conversion [23, 45].

The preliminary results confirm the benefits of using neural voice conversion in the pretraining phase, because data augmentation still preserves both the generality of the model across speakers' genders. Conversely, neural transformations turn out to be harmful in the finetuning phase because model specialization is likely penalized by the presence of artifacts introduced by neural models that lead to the alteration of class-specific characteristic of the vocalisations. Recognizing emotion from audio signals is a challenging task and requires the development of techniques to address the variability of the emotion-related acoustic properties that characterizes the same emotions across speakers.

2 PRIOR WORKS ON AUDIO EMOTION RECOGNITION

Audio Emotion Recognition (AER) commonly entails the following steps: (1) feature extraction from the raw audio, (2) training and application of an emotion classifier on the extracted features. The most commonly used features include, among others, zero-crossing rate, spectral entropy, and chroma vectors [13, 36]. The adopted classifiers span from traditional models, such as Support Vector Machines [36], and Gaussian Mixture Model [39], to Deep Learning models [43]. Recently, a particular attention has been paid to the adoption of Transformers architectures [42]. The developed solutions (e.g., Wav2Vec 2.0 [5], WavLM [7], HuBert [17])

*All authors contributed equally to this research.

Due to their inherent model complexity and flexibility, neural augmentation approaches are likely to be more robust than traditional ones (e.g., pitch shifting). However, they may suffer from the the presence of noisy signals (e.g., the presence of non-speech vocalizations in the input data for AER-NV) or multiple speakers within the same audio track.

4.2 Model pretraining

In our experiments we use the original versions of WavLM [7] and Wav2Vec2.0 [5]. Specifically, we started from the pretrained checkpoints[3] that are learned using self-supervised training objectives.

Prior to specializing them on the proposed task, we adopt an additional pretraining strategy based on contrastive learning. The outcome is latent vector space in which audio tracks expressing the same emotion are represented in a similar way, regardless of the speaker's gender (see Figure 1). To this aim, we train the model using both positive and negative pairs.

Given the anchor audio element, for positive samples we adopt a self-supervised approach to contrastive learning based on data augmentation. The key idea is to leverage data augmentation to make the model more generalizable across different speakers' genders. Specifically, a positive sample corresponds to the same recording augmented using one of the techniques described in the previous section. By altering the original sample related to female vocalizations we build a synthetic version of the corresponding male vocalization, which is included in the representation of positive pair. Negative samples are generated by adopting a supervised approach tailored to emotion recognition. They are picked from the original dataset expressing a different emotion from the anchor. In such a way, the model will automatically learn how to embed samples in a positive pair close in the vector space.

Embedded vectors, either positive or negative, are compared to each other using the cosine similarity. In other words, the contrastive model learns how to minimize the distance between the original audio vocalizations and their corresponding artificial male versions, while keeping female vocalizations that represent different emotions well separated from each other.

We create two distinct pretrained versions of each model, which differ in how positive examples are generated: the former only leverage the traditional data augmentation technique, while the latter exploits also the neural approach. This choice is due to the inability of the voice conversion step to keep the emotion expressed in the audio well defined, therefore we try to remove it from one of the pretraining versions.

4.3 Model finetuning

The following versions of the finetuning step are considered:
(1) No data augmentation at the finetuning stage.
(2) Alter half of the training data samples using pitch shifting and keep the remaining ones unchanged.
(3) Alter half of the training data samples with both pitch shifting and voice conversion technique (in the same proportion).
(4) Alter all the training samples using pitch shifting only.

[3]Wav2Vec 2.0: https://huggingface.co/facebook/wav2vec2-base, WavLM: https://huggingface.co/microsoft/wavlm-base

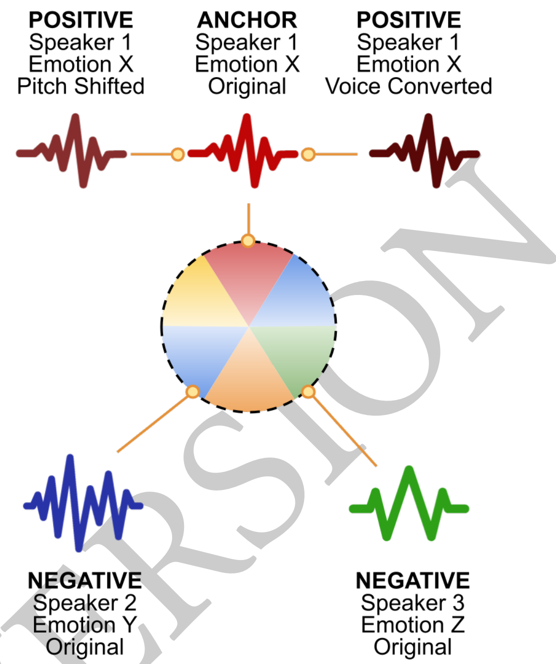

Figure 1: Pretraining strategy based on contrastive learning. Positive samples are an augmented version of the anchor. Negative samples correspond to speakers expressing different emotions.

Table 1: Training time for the pretraining and finetuning steps. The training time is given in seconds per epoch.

Model	Pretraining	Finetuning
Wav2Vec 2.0	12.5s	10s
WavLM	15s	12.5s

Similar to the previous step, data augmentation is aimed at improving the generality of the trained models across different speaker's genders.

5 EXPERIMENTAL RESULTS

5.1 Experimental Design

Hardware settings. The experiments are performed on a machine equipped with AMD® Ryzen 9® 3950X CPU, Nvidia® RTX 3090 GPU, and 128 GB of RAM running Ubuntu 21.10.

Transformers setup. We test four configurations in the pretraining step. We use the Cosine Embedding Loss applied to the embeddings extracted from the last hidden layer of the model. These procedures last for 200 epochs, using a batch size of 16 with and an initial learning rate of 10^{-6} halved every 40 epochs.

In the finetuning step we use a weighted Cross Entropy Loss, for 200 epochs, using a batch size of 16, an initial learning rate of $3 \cdot 10^{-5}$, a warm up ratio of 0.1 and a linear decay.

The pretraining and finetuning times for the considered models, under the reported hardware settings, are reported in Table 1.

Validation. During each pretraining step the model is evaluated on the development set in order to identify the best checkpoint. Next, each checkpoint is finetuned by applying all the previously described combinations of data augmentation. The tested combinations are summarized in Table 2[4]. It reports, for each tested configuration, the Unweighted Average Recall (UAR) [33], which is computed as the mean recall value over all emotion classes.

5.2 Selected runs

Table 2 reports the UAR of the selected runs and of the baseline method released by the ComParE organizers separately for the development and test sets. The configurations evaluated on the test set are a selection of the top-5 most promising settings. Since the development set consists of female vocalizations only whereas the test set contains male vocalizations the performance scores reported in column *Development* of Table 2 are not decisive to shortlist the configurations applied to the test set. The guidelines used to select the best representatives are given below.

- We choose the model that is known to be most suited to non-verbal content analysis (i.e., WavLM [7]).
- To evaluate the impact of the data augmentation phase, we compare the following settings: (1) pitch shifting only (2) pitch shifting combined with neural voice conversion. We separately analyze the above settings on pretraining and finetuning.
- We conduct an ablation study on the percentage of augmented training samples during the finetuning step.

5.3 Results

The main research findings are summarized below.

Comparison with the baseline method on the Development set. The non-augmented model outperforms the baseline method (UAR 45.94 vs. 39.8), whereas all the augmented versions perform as well as or slightly worse than the baseline. The main reason is that data augmentation is instrumental for generalizing the model across different genders. As expected, such a generalization process is beneficial for classifying male vocalizations in the test but is harmful on the female vocalizations in the development set.

Comparison with the baseline method on the Test set. On the test set the UAR scores are all lower than those achieved by the baseline. However, data augmentation has shown to improve the original model performance for all the tested configurations (e.g., UAR 33 with P.S. Pretraining+Finetuning vs. 28.5 with no augmentation).

Comparison between Wav2Vec and WavLM. Wav2Vec 2.0 and WavLM show comparable performance on the development set. However, WavLM is, by construction, more suitable for speaker-related tasks [7]. For this reason, WavLM is the preferred model for the selected runs on the test set.

[4]For the sake of brevity, in the tests we omitted the least interesting combinations.

Effect of data augmentation. The model version without any form of data augmentation achieves the worst performance on the Test set because it has shown to be not robust enough to classify male vocalisation as well. Conversely, the configurations including data augmentation techniques during the pretraining phase are the best performing ones. Augmenting data in the finetuning phase turns out to be not beneficial because the resulting model lacks of a sufficient level of specialization.

Comparison between augmentation techniques. Neural Voice Conversion (V.C.) is less effective than Pitch Shifting (P.S.) on the development Set because the quality of the converted speech is not always satisfactory. Conversely, on the test set the integration of both V.C. and P.S. is beneficial as improves the generality of the model across speakers' genders.

Effect of data augmentation ratio. To further assess the effect of Pitch Shifting, we conduct an experiment by setting the ratio of data augmentation during finetuning to 100% (i.e., the model is trained using only pitch-shifted samples). The results show that it slightly decrease the performance on the test set. This could be expected, since the network only learns with augmented samples, thus it may be difficult for the network to identify the correct class in real data.

6 CONCLUSIONS AND FUTURE WORK

We presented a solution based on Transformers and data augmentation for the Vocalisation sub-challenge of the ComParE 2022 Grand Challenge [35]. We leveraged a contrastive learning approach to achieve the necessary model generality across speakers' genders while mitigating the negative effects on emotion recognition performance. Even though slightly lower than the baseline, the achieved results confirm the expectations about the effect of the data augmentation techniques on the Test performances (i.e., on the male vocalizations) and leave room for several future works. Specifically, we will explore the effect of other augmentation techniques aimed at bringing audio recordings of opposite-sex speakers closer. We also plan to remove codebooks-based quantization from the tested models: it is suitable for spoken content analysis but not necessarily beneficial for non-verbal emotion recognition.

ACKNOWLEDGMENTS

The research leading to these results has been partly funded by the SmartData@PoliTO center for Big Data and Machine Learning technologies.

REFERENCES

[1] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports* 2, 7 (2020), e12189. https://doi.org/10.1002/eng2.12189 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/eng2.12189
[2] Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pre-training for speech recognition. https://doi.org/10.48550/ARXIV.1911.03912
[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. https://doi.org/10.48550/ARXIV.2202.03555
[4] Alexei Baevski, Steffen Schneider, and Michael Auli. 2020. vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rylwJxrYDS

Table 2: Results obtained on Development and Test sets. PS and VC indicate *Pitch Shifting* and *Voice Conversion* respectively. The per-dataset best performer is highlighted in bold ✓* denote the data augmentation steps that are applied to 100% of the training examples.

VOC-C Evaluation						
Pretraining		Finetuning		Development		Test
PS	VC	PS	VC	Wav2Vec2 UAR	WavLM UAR	WavLM UAR
No pretraining				**45.94**	45.04	28.5
✓				38.86	40.05	-
✓		✓		39.32	40.83	33.1
✓		✓	✓	38.33	34.28	-
✓	✓			37.25	36.84	-
✓	✓	✓		38.13	35.83	33.5
✓	✓	✓*		30.97	31.42	33.0
✓	✓	✓	✓	35.47	36.35	31.9
Baseline				39.8		**37.4**

[5] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 12449–12460.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42 (2008), 335–359.

[7] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2021. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. https://doi.org/10.48550/ARXIV.2110.13900

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 1597–1607. https://proceedings.mlr.press/v119/chen20j.html

[9] Donald G. Childers and Kuang chieh Wu. 1991. Gender recognition from speech. Part II: Fine analysis. *The Journal of the Acoustical Society of America* 90 4 Pt 1 (1991), 1841–56.

[10] Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 292–298. https://doi.org/10.1109/ACII.2017.8273615

[11] Yu-An Chung and James Glass. 2019. Generative Pre-Training for Speech with Autoregressive Predictive Coding. https://doi.org/10.48550/ARXIV.1910.12607

[12] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James R. Glass. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. In *INTERSPEECH*.

[13] Stuart Cunningham, Harrison Ridley, Jonathan Weinel, and Richard Picking. 2021. Supervised machine learning for audio emotion recognition. *Personal and Ubiquitous Computing* 25, 4 (2021), 637–650.

[14] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes. 2000. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering* 47, 7 (2000), 829–837. https://doi.org/10.1109/10.846676

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 9726–9735.

[16] Jia-Hao Hsu, Ming-Hsiang Su, Chung-Hsien Wu, and Yi-Hsuan Chen. 2021. Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1675–1686. https://doi.org/10.1109/TASLP.2021.3076364

[17] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[18] Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su, and Yi-Hsuan Chen. 2019. Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5866–5870. https://doi.org/10.1109/ICASSP.2019.8682283

[19] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* 7 (2019), 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

[20] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. 2021. Data Augmenting Contrastive Learning of Speech Representations in the Time Domain. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. 215–222. https://doi.org/10.1109/SLT48900.2021.9383605

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18661–18673. https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf

[22] Tomi Kinnunen, Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, and Zhenhua Ling. 2018. The Voice Conversion Challenge 2018: database and results. (2018).

[23] Jheng-hao Lin, Yist Y Lin, Chung-Ming Chien, and Hung-yi Lee. 2021. S2VC: A Framework for Any-to-Any Voice Conversion with Self-Supervised Pretrained Representations. *arXiv preprint arXiv:2104.02901* (2021).

[24] Benjamin Milde and Chris Biemann. 2018. Unspeech: Unsupervised Speech Context Embeddings. *Proc. Interspeech 2018* (2018), 2693–2697.

[25] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-Supervised Speech Representation Learning: A Review. https://doi.org/10.48550/ARXIV.2205.10643

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. https://doi.org/10.48550/ARXIV.1807.03748

[27] Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. 2019. Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. *Proc. Interspeech 2019* (2019), 161–165.

[28] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. https://doi.org/10.48550/ARXIV.2104.03502

[29] Cyril Pernet and Pascal Belin. 2012. The Role of Pitch and Timbre in Voice Gender Categorization. *Frontiers in Psychology* 3 (2012). https://doi.org/10.3389/fpsyg.2012.00023

[30] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for Robust Speech Recognition. https://doi.org/10.48550/ARXIV.2001.09239

[31] Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive Learning of General-Purpose Audio Representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3875–3879. https://doi.org/10.1109/ICASSP39728.2021.9413528

[32] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. *Proc. Interspeech 2019* (2019), 3465–3469.

[33] Bjorn Schuller and Anton Batliner. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing* (1st ed.). Wiley Publishing.

[34] B. Schuller, G. Rigoll, and M. Lang. 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. I–577. https://doi.org/10.1109/ICASSP.2004.1326051

[35] Bjorn W. Schuller, Anton Batliner, Shahin Amiriparian, Christian Bergler, Maurice Gerczuk, Natalie Holz, Pauline Larrouy-Maestri, Sebastian P. Bayerl, Korbinian Riedhammer, Adria Mallol-Ragolta, Maria Pateraki, Harry Coppock, Ivan Kiskin, Marianne Sinka, and Stephen Roberts. 2022. The ACM Multimedia 2022 Computational Paralinguistics Challenge: Vocalisations, Stuttering, Activity, & Mosquitos. In *Proceedings ACM Multimedia 2022*. ISCA, Lisbon, Portugal. to appear.

[36] MS Sinith, E Aswathi, TM Deepa, CP Shameema, and Shiny Rajan. 2015. Emotion recognition from audio signals using Support Vector Machine. In *2015 IEEE recent advances in intelligent computational systems (RAICS)*. IEEE, 139–144.

[37] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. 2020. CURL: Contrastive Unsupervised Representations for Reinforcement Learning. https://doi.org/10.48550/ARXIV.2004.04136

[38] Ingo R. Titze and Daniel w. Martin. 1994. Principles of voice production.

[39] Dinh-Son Tran, Hyung-Jeong Yang, Soo-Hyung Kim, Guee Sang Lee, Luu-Ngoc Do, Ngoc-Huynh Ho, and Van Quan Nguyen. 2018. Audio-based emotion recognition using GMM supervector an SVM linear kernel. In *Proceedings of the 2nd*

International Conference on Machine Learning and Soft Computing. 169–173.

[40] Wei-Cheng Tseng, Chien-yu Huang, Wei-Tsung Kao, Yist Y. Lin, and Hung-yi Lee. 2021. Utilizing Self-Supervised Representations for MOS Prediction. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek (Eds.). ISCA, 2781–2785. https://doi.org/10.21437/Interspeech.2021-2013

[41] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *NIPS*.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[43] Detai Xin, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Exploring the Effectiveness of Self-supervised Learning and Classifier Chains in Emotion Recognition of Nonverbal Vocalizations. *arXiv preprint arXiv:2206.10695* (2022).

[44] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing Universal PERformance Benchmark. https://doi.org/10.48550/ARXIV.2105.01051

[45] Assila Yousuf and David Solomon George. 2022. Parallel data free singing voice conversion with cycle-consistent BEGAN. *Materials Today: Proceedings* 58 (2022), 157–161. https://doi.org/10.1016/j.matpr.2022.01.169 International Conference on Artificial Intelligence & Energy Systems.