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

Lorenzo Vaiani\textsuperscript{*}  
lorenzo.vaiani@polito.it  
Politecnico di Torino  
Turin, Italy

Alkis Koudounas\textsuperscript{*}  
aliks.koudounas@polito.it  
Politecnico di Torino  
Turin, Italy

Moreno La Quatra\textsuperscript{*}  
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  
elen.baralis@polito.it  
Politecnico di Torino  
Turin, Italy

\textsuperscript{*}All authors contributed equally to this research.

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 it is harmful at the finetuning stage as hinders model specialization for the task of non-verbal emotion recognition.

CSC CONCEPTS

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 of 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 Computational Paralinguistic Challenge (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])...
have achieved substantial performance improvements, on speech-related tasks, against traditional techniques on benchmark data (e.g., SUPERB [44], IEMOCAP [6], and RAVDESS [28]). Nonverbal Vocalization [16, 18] is a specific AER subtask (hereafter denoted by AER-NV), which has already been addressed in the context of multimodal learning [10].

3 AUDIO SPEECH CLASSIFICATION VIA SELF-SUPERVISED LEARNING

The advent of Deep Learning techniques has radically changed the ways of processing and classifying audio speech data. Since labeling data is a labour-intensive task, a huge body of work has been devoted to learning speech data representations using self-supervised learning [25]. Within this scope, a pre-trained representation model, also called “upstream” model, is learnt first. Then, the model is fine-tuned to tailor the representation to a specific downstream task (e.g., Speech Emotion Recognition). To learn upstream models, the following tasks have been addressed:

1. generative (e.g., VQ-VAE [41], APC [11, 12], PASE [27] and PASE* [30]),
2. predictive (e.g., DiscreteBERT [2], HuBERT [17], WavLM [7] and Data2Vec [3]),
3. contrastive learning (e.g., Contrastive Predictive Coding [26], Unspeech [24], Wav2Vec [32], VQ-Wav2Vec [4] and Wav2Vec2.0 [5]).

This work addresses the use of contrastive learning to pre-train a model suited to AER-NV using self-supervised learning.

3.1 Contrastive Learning

Contrastive learning has achieved state-of-the-art performance in several application contexts, among which computer vision [8, 15] and reinforcement learning [37]. Recently, it has been used to self-learn acoustic data representations [20, 31]. The key idea is to self-learn the key data characteristics by letting the neural model learn how to map similar examples and to discriminate dissimilar ones. Given a data point in the original dataset, namely the anchor, it is paired with an altered version of itself to generate a positive pair. Data alterations are typically obtained via data augmentation. When the input data is labeled, positive pairs may consist of points belonging to the same class [21]. Alternatively, they can be audio fragments belonging to the same audio track [31]. However, in many real-world application scenarios, such as the ComParE task, the presence of bias in the annotated data may limit the model portability towards different contexts. Specifically, in ComParE the annotated audio tracks are all related to female vocalizations. Hence, learning predictive patterns from positive pairs does not preserve the generality across speakers’ genders.

An alternative contrastive learning approach tailored to audio speech data consists in leveraging data augmentation techniques, such as pitch shifting, to improve the robustness of the pretrained model [20]. The latter approach can be deemed as helpful for mitigating the gender bias in the source data, for instance, by considering the established role of pitch and timbre in voice gender categorization [29].

To address ComParE we adopt a mixed contrastive approach relying on both self-supervised and supervised learning. Specifically, to generate positive pairs we rely on augmented data whereas negative pairs are determined according to a combination of samples belonging to the different emotion classes.

4 THE PROPOSED METHOD

The method proposed for the Computational Paralinguistic Challenge (ComParE) [35] consists of:

- a data augmentation step, which produces altered audio samples that can be used to build the contrastive data pairs (see Section 4.1).
- a model pretraining step, in which the Transformers architecture learns how to solve an upstream, more general task via contrastive learning (see Section 4.2).
- a model finetuning step, in which the pretrained model is specialized for the AER-NV downstream task (see Section 4.3).

The pretraining and finetuning steps are complementary to (i) learning gender-unbiased audio representations and (ii) solving the classification task, respectively. The project source code is available for research purposes.

4.1 Data Augmentation

We explore the use of both traditional and neural approaches.

4.1.1 Traditional approach: pitch shifting. As discussed in [38], there exists an evident sexual dimorphism between the vocal apparatus of male and female adults. This causes the main dissimilarities we hear while listening to female and male voices, in particular for what concerns the mean fundamental frequency of phonation (F0) and the formant frequencies [9]. The fundamental frequency (related to the perceived pitch) is generally inversely proportional to the size of the source [29]. This means that adult males present voices with a lower pitch with respect to adult females.

We leverage pitch shifting techniques to augment data by lowering the pitch of female audio speeches to simulate male voices.

4.1.2 Neural approach: automatic voice conversion. We adopt neural network models to automatically convert female voices to male one in order to generate the augmented data samples while mitigating the gender bias in the input data.

We rely on a self-supervised Voice Conversion (VC) model [23] that is able to transform the identity of a given source audio to that of a different target audio. By leveraging the training split of the data collection released for the Voice Conversion Challenge (VCC 2018) [22] and the open-source implementation of the S2VC framework, we use as source voices the female speakers of the VCC dataset and each of the male speakers as target.

The model is trained on the speech recordings of the VCC 2018 dataset [40] and the resulting voice conversion model is used to augment the training samples of the vocalisation dataset. For each record in the vocalisation dataset that corresponds to a female speaker, its augmented version is generated by using as target voice one random speaker sampled from the VCC dataset.

1https://github.com/VaianiLorenzo/compare2022_vocalisation (Latest access: June 2022)
2https://github.com/howard1337/S2VC (Latest access: June 2022)
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 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 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 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.

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 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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Pretraining</th>
<th>Finetuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wav2Vec 2.0</td>
<td>12.5s</td>
<td>10s</td>
</tr>
<tr>
<td>WavLM</td>
<td>15s</td>
<td>12.5s</td>
</tr>
</tbody>
</table>

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

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
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 the training data.

<table>
<thead>
<tr>
<th>PS VC</th>
<th>PS VC</th>
<th>Wav2Vec2 UAR</th>
<th>WavLM UAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pretraining</td>
<td>✓</td>
<td>✓</td>
<td>45.94</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>38.86</td>
<td>40.05</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>39.32</td>
<td>40.83</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>38.33</td>
<td>34.28</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>37.25</td>
<td>36.84</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>38.13</td>
<td>35.83</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>30.97</td>
<td>31.42</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>35.47</td>
<td>36.35</td>
</tr>
</tbody>
</table>

Baseline 39.8 37.4

References:


