Grósz, Tamás; Porjazovski, Dejan; Getman, Yaroslav; Kadiri, Sudarsana; Kurimo, Mikko

Wav2vec2-based Paralinguistic Systems to Recognise Vocalised Emotions and Stuttering

Published in:
Proceedings of the 30th ACM International Conference on Multimedia

DOI:
10.1145/3503161.3551572

Published: 01/10/2022

Document Version
Publisher's PDF, also known as Version of record

Wav2vec2-based Paralinguistic Systems to Recognise Vocalised Emotions and Stuttering

Tamás Grósz*
tamas.grosz@aalto.fi
Aalto University
Espoo, Finland

Dejan Porjazovski*
dejan.porjazovski@aalto.fi
Aalto University
Espoo, Finland

Yaroslav Getman*
yaroslav.getman@aalto.fi
Aalto University
Espoo, Finland

Sudarsana Reddy Kadiri
sudarsana.kadiri@aalto.fi
Aalto University
Espoo, Finland

Mikko Kurimo
mikko.kurimo@aalto.fi
Aalto University
Espoo, Finland

ABSTRACT

With the rapid advancement in automatic speech recognition and natural language understanding, a complementary field (paralinguistics) emerged, focusing on the non-verbal content of speech. The ACM Multimedia 2022 Computational Paralinguistics Challenge introduced several exciting tasks of this field. In this work, we focus on tackling two Sub-Challenges using modern, pre-trained models called wav2vec2. Our experimental results demonstrated that wav2vec2 is an excellent tool for detecting the emotions behind vocalisations and recognising different types of stutterings. Albeit they achieve outstanding results on their own, our results demonstrated that wav2vec2-based systems could be further improved by ensembling them with other models. Our best systems outperformed the competition baselines by a considerable margin, achieving an unweighted average recall of 44.0 (absolute improvement of 6.6% over baseline) on the Vocalisation Sub-Challenge and 62.1 (absolute improvement of 21.7% over baseline) on the Stuttering Sub-Challenge.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*; • **Information systems** → **Multimedia and multimodal retrieval**.

KEYWORDS

Computational Paralinguistics, Vocalisations, Stuttering, wav2vec2, Data Augmentation, Challenge

ACM Reference Format:
Tamás Grósz, Dejan Porjazovski, Yaroslav Getman, Sudarsana Reddy Kadiri, and Mikko Kurimo. 2022. Wav2vec2-based Paralinguistic Systems to Recognise Vocalised Emotions and Stuttering. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3503161.3551572

*All three authors contributed equally to this research.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3551572

1 INTRODUCTION

With the advancement of automatic speech recognition (ASR) and synthesis models, paralinguistic solutions become ever more important. The field of paralinguistics focuses on recognising *how* something was said instead of *what* was said. This year the ACM Multimedia Computational Paralinguistics Challenge introduced several interesting tasks [16]. This work focuses on tackling two of the sub-challenges, namely, the Vocalisation and Stuttering tasks by employing various modern techniques and adapting them to the special needs of each challenge.

The Vocalisation (VOC-C) Sub-Challenge seeks an answer to the question of how well can we recognise emotions from non-verbal vocal expressions. The added difficulty is the considerable mismatch between the released training (female speakers) and test (male speakers) data, provided by the Variably Intense Vocalizations of Affect and Emotion (VIVAE) Corpus [13, 14]. While the model needs to be trained on female vocalisations, during evaluation, it will receive sounds from male vocalisations. The approaches developed for this task could be valuable for building more accurate spoken emotion recognition systems.

In the Stuttering Sub-Challenge, a selected part of the Kassel State of Fluency corpus (KSF-C) [5, 6] is given to the participants. The aim is to develop solutions that can monitor the speech of stuttering people, and provide feedback about the types of disfluencies recognised in their speech. Such systems would be excellent tools for speech therapy and computer-assisted pronunciation training.

Pre-trained models attracted considerable attention in the field of paralinguistics as most datasets are relatively small compared to speech recognition datasets. The existing solutions employ a wide range of approaches. One common technique is to use the pre-trained systems as feature extractors before training simple classifiers like Support Vector Machines (SVM), an example of this being the baseline solutions using DeepSpectrum [3]. Another popular approach is to pre-train an autoencoder using the audio files and use its embeddings as input for a classifier [11]. Lastly, one can use pre-trained models and fine-tune them on the small paralinguistic data, for example in [17], various image classifier networks were adapted to detect mask-wearing from speech.

In this work, we follow this trend and rely on pre-trained models. In contrast with previous systems, we do not employ networks pre-trained as image classifiers but rather use state-of-the-art wav2vec2

models [4]. These wav2vec2 models are pre-trained in an unsupervised manner using large audio datasets, and after some fine-tuning on a limited supervised set they were shown to achieve excellent performance. Some works were already done to employ these networks on emotion recognition tasks, showing that they have potential uses in paralinguistics [15]. Our empirical results confirm that wav2vec2 is a viable tool for paralinguistic problems, but we need to be careful in selecting the right base model for fine-tuning. Furthermore, we found that despite their good performance, they are far from being the perfect tool. Combining them with other models results in even more accurate systems.

2 METHODS

2.1 wav2vec2 audio classifier

In recent years, pre-trained solutions have become popular for paralinguistic tasks, and we could observe a significant shift toward end2end neural solutions [8]. A considerable shortcoming of most pre-trained approaches is that they employ unrelated, mostly image classifier models. To alleviate this, we test if fine-tuning the state-of-the-art wav2vec2 models could lead to better results [4]. These models are generally pre-trained on a large unsupervised audio dataset and fine-tuned on a small set to perform ASR, excelling in low-resource scenarios [2, 19] and on out-of-domain data [20]. In this work, we experimented with several models shared on HuggingFace and use them as audio classifiers.

To fine-tune the wav2vec2 models, we used the Sequence Classification interface, which simply connects a new output layer to the averaged pooled context vector as explained in [9]. Only the feature extractor was frozen during training (in most cases) which allowed the model to adjust its transformer and, consequently, the context vectors for the given task. Our previous experiences on pronunciation evaluation tasks confirmed that this approach is better than using the wav2vec2 as a static feature extractor [10]. All wav2vec2 models were fine-tuned using one GPU (NVIDIA RTX 2080 Ti or Tesla V100), with a batch size of 8, and the initial learning rate was set to 3e-5. For more technical details about our systems see our code repository [1].

2.2 VGGish + TCN audio classifier

The large pre-trained models serve as powerful feature extractors, which with small amount of fine-tuning data can be adapted to a broad range of tasks. The development of such pre-trained models using large-scale audio data has allowed for extraction of more robust audio features. One such architecture is the VGGish [12], trained on large amount of YouTube videos, which later became the YouTube-8M dataset [1]. The VGGish features have been successfully incorporated in various audio classification tasks, such as sound event detection [7, 18], showcasing the benefit of large-scale pre-training.

Even though the VGGish embeddings can be directly used for classification, we further processed them using a Temporal Convolutional Network (TCN) [11], consisting of convolutional blocks with dilated connections. This way, the embeddings are more adapted for the specific task. The adaptation is done by first extracting VGGish embeddings from each 1 second interval and then processing those embeddings using the TCN. In the cases where the audio length is less than a second, the audio is zero-padded to reach 1 second length.

2.3 Data augmentation

The dataset for vocalisation task (6 class problem) consists of around 625 training samples, falling into the low-resource category. Furthermore, the mismatch between training (6 female speakers) and testing data (2 male speakers) makes the task even more challenging. To reduce the mismatch between training and testing splits, we artificially created male versions of the training samples by lowering the fundamental frequency. This doubled our training data, resulting in 1250 samples. The dataset for stuttering task consists of 4601 speech segments (3 seconds length) from 37 German speakers, containing 7 categories of stuttering and some non-disfluent samples. More details about the datasets can be found in [16].

3 EXPERIMENTS AND RESULTS

3.1 Vocalisation

For the Vocalisation task, five different wav2vec2 models were investigated. The first one was a pre-trained medium-sized (300M parameters) network ($wav2vec2_M$ *no ASR*). In order to see if the intermediate fine-tuning helps, we also selected a multilingual ASR version of the first one ($wav2vec2^*_M$). Furthermore, two additional models were selected, multilingually pre-trained wav2vec2 solutions fine-tuned for German ASR ($wav2vec2^{de}_M$ and $wav2vec2^{de}_L$). The selection of these two models was motivated by the fact that these models use data containing vocalisations from German speakers [16]. The main difference between the two German models is their size, the $wav2vec2^{de}_L$ containing three times more parameters than $wav2vec2^{de}_M$. Lastly, we also experimented with applying a wav2vec2 fine-tuned for emotion recognition ($wav2vec2^{er}_M$). More details about these models can be found in our codes [1].

On the non-augmented data, it is observed that the models fine-tuned for ASR performed better than the raw pre-trained models, see table 1 (development dataset). Furthermore, it can be seen that the German ASR models yielded considerably better results than the multilingual models. The best result is achieved by the model that was trained for emotion recognition before being fine-tuned on this data. These observations highlight the importance of selecting the most suitable model for the task.

Besides the pre-trained wav2vec2 models, we additionally experimented with the pre-trained VGGish model as a feature-extractor. Initial investigations revealed that VGGish features alone gave poor performance on the development set. Due to that, we additionally processed the embeddings using the TCN (model 6 in table). By utilising only the non-augmented data, we observed that the system is falling behind all the wav2vec2 models by a significant margin, see table 1. By applying the augmentation techniques described in section 2.3, we observed a notable improvement, bringing this approach closer to the wav2vec2 solutions. To save time, we only trained $wav2vec2^{de}_M$ on the augmented data, but it did not show an improvement. A possible explanation could be that the large wav2vec2 was able to overfit on the augmented data and thus lost

[1] https://github.com/aalto-speech/ComParE2022

Table 1: Development set (dev) UARs (unweighted average recall) on the Vocalisation task.

Model	Orig.	Male-like	Augmented
1. $wav2vec2_M$ no ASR	35.5	–	–
2. $wav2vec2^*_M$	37.3	–	–
3. $wav2vec2^{de}_M$	39.5	37.6	38.9
4. $wav2vec2^{de}_L$	39.4	–	–
5. $wav2vec2^{er}_M$	**41.3**	–	–
6. vggish+TCN	32.1	35.6	35.0
baseline (BoAWs)	39.6	–	–

Table 2: Final dev and test UARs on the Vocalisation task. The test results are the performances of re-trained models on all labelled data (merged train and dev sets).

Model	Dev	Test
$wav2vec2^{er}_M$	41.3	34.5
combination of 3 and 6	42.0	**44.0**
combination of 5 and 6	43.9	38.0
combination of 3, 5 and 6	**46.2**	41.2
baseline (BoAWs)	39.6	37.4

performance on the original female development set. Further, the analysis of the predicted labels from both $wav2vec2^{de}_M$ and VGGish+TCN models showed that the system trained on the whole augmented data matches more closely the true labels class distribution on 4 out of 6 classes. Due to that, in the rest of the paper experiments are carried out with the models trained on the whole augmented data.

Table 2 shows the results of the submitted systems for the test set (along with results on dev set). First of all, we selected the best single non-augmented model ($wav2vec2^{er}_M$) (see table 1), which proved to have overfitted to the female voices in the training and dev sets and had a below-baseline performance on the test set. To avoid such issues, in the other submissions, we opted to combine the models with the augmented VGGish+TCN. On the development set, we observed considerable improvements due to this combination, however, the same trend is not true for the test results. The performance of the ensembles containing $wav2vec2^{er}_M$ (indicated as *combination of* 5 *and* 6) proved to be poorer than just combining the monolingual $wav2vec2^{de}_M$ and the augmented VGGish+TCN (*combination of 3 and 6*). Overall we can see that ensembling with the augmented model (i.e., *combination of* 3, 5 *and* 6 in the table) leads to the best performance on the development set, but not on the test set due to the gender mismatch. The best system (i.e., *combination of* 3 *and* 6) outperformed the baseline on the test set by a considerable margin, achieving an UAR of 44.0, an absolute improvement of 6.6%.

3.2 Stuttering

In the Stuttering Sub-Challenge, we employed the same monolingual $wav2vec2^{de}_M$ model, which proved to be the best for the Vocalisation task. In addition, we experimented with a smaller model ($wav2vec2^{de}_S$) pre-trained and ASR fine-tuned purely on German speech, containing three times less parameters compared to $wav2vec2_M$. These models outperformed the baseline solutions, as can be seen in table 3 (for the dev set). Next, we turned our attention to model size and training different parts of the architecture. We observed having more trainable parameters is important, the largest model ($wav2vec2^{de}_L$) even benefited from training the CNN part (*full train*) at the expense of the training times (UAR of 59.3 with an absolute improvement of 31.2% over baseline). The most important model part proved to be the Transformer, freezing its weights yielded extremely bad results. In addition, the model pre-trained solely on German speech ($wav2vec2^{de}_S$) provided competitive performance (UAR of 51.2, an absolute improvement of 23.1% over baseline) regardless of the lower number of parameters and amount of pre-training data compared to other wav2vec2 systems.

Upon closer inspection of the predicted classes, we observed that most models struggle to detect certain classes. The two least recognised classes are word repetitions and garbage sentences, the two rarest classes accounting for only 2% of the training data each. Regardless of the rarity of these disfluencies, their low recognition rate is somewhat expected since the averaged pooled embeddings do not necessarily contain any high-level linguistic information needed for the detection of these phenomena. This observation motivated us to explore other solutions, which process texts hoping that such a system would achieve higher accuracy.

As an initial step, we used the $wav2vec2^{de}_M$ system to transcribe the recordings and processed the ASR transcript with numerous models. Although we tried various methods and features on the ASR transcripts, in the end, a straightforward solution yielded the best results. The main weakness of most systems was that they struggled to differentiate between garbage and non-disfluent samples. We observed that the length of the ASR transcripts greatly differ for these categories. Our solution exploits this fact to re-classify some of the audios recognised as non-disfluent and re-classify them as garbage if the length of their transcript was below a threshold (determined using the transcripts of training data).

Table 4 shows the performance of the final models for the dev and test sets of Stuttering challenge. From the table, it can be observed that the big wav2vec2 model significantly outperformed the medium one on both sets. With the additional post-processing of replacing some of the non-disfluent labels with garbage, based on the transcript length, we gained further improvement, achieving an UAR of 62.1 and 61.3, an absolute improvement of 21.7% and 33.2% over baseline on test and dev sets, respectively. Lastly, we tried various ensembles of the two wav2vec2 models and the TCN network. Unfortunately, none of the ensembles managed to outperform the best model, signalling that the text-based corrections are more vital for this task, and they require further investigation.

4 CONCLUSIONS

In this work, we presented wav2vec2-based paralinguistic solutions for two sub-tasks of the ACM Multimedia Computational Paralinguistics Challenge, namely the Stuttering and Vocalisation Sub-Challenges. Our results demonstrated that carefully pre-trained wav2vec2 models are superior alternatives to models pre-trained

Table 3: Dev UARs on the Stuttering task.

Model	input	Dev UAR
TCN	vggish	26.4
$wav2vec2_S^{de}$	audio	51.2
$wav2vec2_M^{de}$	audio	50.1
$wav2vec2_L^{de}$	audio	54.5
$wav2vec2_L^{de}$ full train	audio	**59.3**
$wav2vec2_L^{de}$ frozen Transf.	audio	14.3
baseline	DeepSpect.	28.1

Table 4: Final dev and test UARs on the Stuttering task.

Model	Dev	Test
$wav2vec2_M^{de}$	50.1	57.1
$wav2vec2_L^{de}$ full train	59.3	61.5
+ text-based correction	**61.3**	**62.1**
$wav2vec2_L^{de}$ full train + TCN	58.7	–
+ text-based correction	61.1	61.9
ensemble of all 3	59.7	–
+ text-based correction	60.7	–
baseline (DeepSpectrum)	28.1	40.4

on image classification tasks. Furthermore, we demonstrated the importance of choosing the right model for the task at hand. We showed that it is beneficial to leverage a model fine-tuned for a similar downstream task or on a dataset from a close domain rather than a raw pre-trained one. We found that performing an ASR finetuning before the paralinguistic one is highly beneficial. On the monolingual tasks, model size does matter when pre-trained on the same acoustic data. Also, models initially tuned for monolingual ASR achieved higher scores, and we saw that a smaller model pre-trained on monolingual data is able to provide competitive performance compared to larger systems pre-trained on multilingual data.

In addition, it was found that adjusting the Transformer weights for the target downstream task is very crucial. In contrast, updating the weights of the feature encoder does not always guarantee to improve the results. The effect of fine-tuning the CNN weights might be influenced by several factors, such as overall size of the model, amount of training data, as well as how far target data is from the domain of pretraining data. The best wav2vec2 solutions outperformed the baseline systems, but they are not perfect tools; combining them with other models yielded further improvements.

ACKNOWLEDGMENTS

The computational resources to perform the experiments were provided by Aalto ScienceIT. This work was supported by NordForsk through the funding to Technology-enhanced foreign and second-language learning of Nordic languages, project number 103893. We are grateful for the Academy of Finland project funding number 345790 in ICT 2023 programme's project "Understanding speech and scene with ears and eyes".

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Ragheb Al-Ghezi, Yaroslav Getman, Aku Rouhe, Raili Hildén, and Mikko Kurimo. 2021. Self-Supervised End-to-End ASR for Low Resource L2 Swedish. In *Proc. Interspeech 2021*. 1429–1433. https://doi.org/10.21437/Interspeech.2021-1710

[3] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. Snore Sound Classification Using Image-Based Deep Spectrum Features. In *Interspeech 2017*. ISCA, 3512–3516.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. 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., 12449–12460. https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[5] Sebastian P Bayerl, Florian Hönig, Joëlle Reister, and Korbinian Riedhammer. 2020. Towards automated assessment of stuttering and stuttering therapy. In *International Conference on Text, Speech, and Dialogue*. Springer, 386–396.

[6] Sebastian P Bayerl, Alexander Wolff von Gudenberg, Florian Hönig, Elmar Nöth, and Korbinian Riedhammer. 2022. KSoF: The Kassel State of Fluency Dataset–A Therapy Centered Dataset of Stuttering. *arXiv preprint arXiv:2203.05383* (2022).

[7] Gianmarco Cerutti, Rahul Prasad, Alessio Brutti, and Elisabetta Farella. 2019. Neural Network Distillation on IoT Platforms for Sound Event Detection.. In *Interspeech*. 3609–3613.

[8] Harry Coppock, Alex Gaskell, Panagiotis Tzirakis, Alice Baird, Lyn Jones, and Björn Schuller. 2021. End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study. *BMJ Innovations* 7, 2 (2021), 356–362. https://doi.org/10.1136/bmjinnov-2021-000668 arXiv:https://innovations.bmj.com/content/7/2/356.full.pdf

[9] Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2021. Exploring wav2vec 2.0 on Speaker Verification and Language Identification. In *Proc. Interspeech 2021*. 1509–1513. https://doi.org/10.21437/Interspeech.2021-1280

[10] Yaroslav Getman, Ragheb Al-Ghezi, Ekaterina Voskoboinik, Tamás Grósz, Mikko Kurimo, Giampiero Salvi, Torbjørn Svendsen, and Sofia Strömbergsson. 2022. wav2vec2-based Speech Rating System for Children with Speech Sound Disorder. In *Proc. Interspeech 2022*. to appear.

[11] Albert Haque, Michelle Guo, Prateek Verma, and Li Fei-Fei. 2019. Audio-linguistic embeddings for spoken sentences. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7355–7359.

[12] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[13] Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. 2021. The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific reports* 11, 1 (2021), 1–10.

[14] Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. 2022. The variably intense vocalizations of affect and emotion (VIVAE) corpus prompts new perspective on nonspeech perception. *Emotion* 22, 1 (2022), 213.

[15] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In *Proc. Interspeech 2021*. 3400–3404. https://doi.org/10.21437/Interspeech.2021-703

[16] Björn 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.

[17] Jeno Szep and Salim Hariri. 2020. Paralinguistic Classification of Mask Wearing by Image Classifiers and Fusion. In *Proc. Interspeech 2020*. 2087–2091. https://doi.org/10.21437/Interspeech.2020-2857

[18] Huang Xie and Tuomas Virtanen. 2021. Zero-shot audio classification via semantic embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1233–1242.

[19] Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2020. Applying wav2vec2.0 to Speech Recognition in various low-resource languages. *ArXiv* abs/2012.12121 (2020).

[20] Juan Zuluaga-Gomez, Amrutha Prasad, Iuliia Nigmatulina, Saeed Sarfjoo, Petr Motlicek, Matthias Kleinert, Hartmut Helmke, Oliver Ohneiser, and Qingran Zhan. 2022. How Does Pre-trained Wav2Vec2. 0 Perform on Domain Shifted ASR? An Extensive Benchmark on Air Traffic Control Communications. *arXiv preprint arXiv:2203.16822* (2022).