

THE VOICE OF AN INSTRUMENT: ANALYSIS OF X-VECTORS FOR MUSIC EMOTION RECOGNITION

Mariana Rodríguez Castañeda
Universidad Nacional Autónoma de México
mariana.rodriguez.c@comunidad.unam.mx

Iran R. Roman
Queen Mary University of London, London, UK
i.roman@qmul.ac.uk

ABSTRACT

This late-breaking demo details our reproduction of the experiment discussed in the paper by Aldene & Provost (2023) and its adaptation to musical excerpts. Their research originally explores how x-vectors constructed from speech utterances capture emotional information. We first present our findings based on the data and setup used in the original paper. Subsequently, we conduct a similar experiment utilizing a selection of musical excerpts to find similar patterns of results across both experiments.

1. INTRODUCTION

It has been proven that certain emotions can significantly improve productivity in workplaces and learning in schools [1, 2], while others can lead to trigger disorders such as depression [3, 4]. In audio processing (and more widely in machine learning), feature selection is a critical step in order to carry out automatic emotion recognition [5]. Recently, Aldene & Provost showed that x-vectors, the embeddings of a deep neural network for speaker identification [6], robustly capture paralinguistic features [7]. Their experiments provide empirical evidence that the emotional content in speech utterances is indeed embedded in x-vectors. In this late breaking demo, we hypothesized that, similar to speech, the emotions conveyed by musical excerpts could also be captured by x-vectors. Our preliminary experiments replicate the findings of Aldene & Provost [7], and we also applied the same analysis to x-vectors obtained from musical excerpts. Our results reveal a pattern similar to those observed by Aldene & Provost in their speech studies. [7].

2. METHODOLOGY

2.1 Datasets

For the reproduction of results by Aldene & Provost, we used the VESUS dataset, which contains over 250 distinct phrases, each read by ten actors in five emotional states [8].

It also provides a crowd-sourced human rating for the perceived emotional content of each utterance [8]. We used the utterances that were rated to be consistent with the emotion intended to convey.

To carry out experiments with music, we created a dataset of musical excerpts. Neutral excerpts were designed to contain scales and sequences of single notes. We created excerpts with violin, piano, and guitar tones from the Freesound technical demo dataset [9]. Bass guitar neutral excerpts were generated using the IDMT-SMT-Bass-Single-Track Dataset [10], which contains single recorded notes with ten different bass-related playing techniques. We generated 150 neutral tracks for each instrument, which were split into training, validation, and testing sets (60%, 20%, and 20%, respectively). Each track was split into two-second-long clips, matching the duration of utterances used by Aldeneh & Provost [7]

The emotional musical excerpts were obtained from the MTG-Jamendo Dataset, which contains over 55,000 audio tracks with 195 tags from genre, instrument, and mood/theme categories [11]. To obtain the highest number of 'angry', 'happy', and 'sad' excerpts, tracks labeled as 'energetic', 'dark', 'heavy', 'horror', 'powerful', 'action', and 'fast' were considered as 'angry'. Tracks labeled as 'happy', 'fun', 'funny', 'uplifting', 'upbeat', 'positive', 'motivational', 'children', and 'Christmas' were considered as 'happy'. And tracks labeled as 'meditative', 'slow', 'sad', 'emotional', 'melancholic', 'deep', 'dream' were considered to be 'sad'. The piano, violin, guitar, and bass were selected for having the highest available audios with these labels. In the end, 30 two-second-long clips for each emotion and instrument were used to match the proportion and durations of the "neutral" test set.

2.2 X-vectors

The x-vector model is a time-delay neural network trained on VoxCeleb1 [12] and VoxCeleb2 [13], datasets with more than 2,000 hours of speech (more than 1 million utterances) from more than 7,000 speaker identities [12, 14]. X-vectors are extracted at layer six of the standard architecture before the non-linearity. This embedding was originally used for spoken language recognition [15] but was then demonstrated to embed emotion and style information [16, 17]. We encoded each audio clip in our datasets with this model, leaving us with the 512-dimensional vector of layer six to do further experiments.



© M. Rodríguez, and I. Roman. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** M. Rodríguez, and I. Roman, "The Voice of an Instrument: Analysis of x-vectors for music emotion recognition", in *Extended Abstracts for the Late-Breaking Demo Session of the 25th Int. Society for Music Information Retrieval Conf.*, San Francisco, United States, 2024.

2.3 Autoencoders

The autoencoder model proposed by Aldene & Provost takes x-vectors as input and consists of an encoder with five down-sampling layers, followed by a decoder with five up-sampling layers, all fully connected. Each layer uses a Tanh activation function, except the final layer. This architecture effectively compresses x-vectors from 512 to 16 dimensions and reconstructs them back to 512 dimensions. We pre-trained the model on the Librispeech dataset, a corpus derived from audiobooks that are part of the LibriVox project, and contains 1000 hours of speech [18]. Training ceased after no improvement in validation loss was observed for five consecutive epochs, using the Adam optimizer with a fixed batch size of 100.

2.4 Experiments

First we fine-tune the autoencoder to optimally reconstruct the neutral speech clips according to cross-validated training. Using this model we reconstruct the neutral test set of speech x-vectors to obtain a baseline reconstruction error. Next, we reconstruct the x-vectors of test set clips with emotional speech. We repeat the same procedure with the musical excerpts data¹. The hypothesis is that if the reconstruction error for the test set neutral clips is smaller than the emotion clips, then this drop in performance is caused by the x-vectors' ability to capture emotional content in the audio signals. In other words, x-vectors serve as evidence that emotion indeed modulates the sound of speech or a musical excerpt.

3. RESULTS & DISCUSSION

Figure 1 presents our reproduction of results by Aldene & Provost. The neutral box plot presents the baseline reconstruction error, whereas the box plots for angry, happy, and sad utterances show an increase compared to the baseline. This indicates that the x-vector captures not only the characteristics of the speaker's voice but also emotional information. In essence, since emotions have an impact on the modulation of a speaker's voice, they also play a role in the x-vector analysis. [7]. On the other hand, Figure 2 demonstrates that adapting Aldene & Provost approach to musical excerpts yields similar trends. The baseline error for musical excerpts with neutral emotion is notably lower than the error obtained when reconstructing excerpts that convey anger, happiness, or sadness.

We want to highlight the concern that, since we sourced musical signals from different datasets, our preliminary results in Figure 2 may be affected by a domain shift effect. Aldene & Provost (and our replication of their results) used a single dataset for both neutral and emotional speech data. In our approach, creating the dataset of musical excerpts using existing datasets required us to sourced from various origins. This important distinction from the original work should be examined in more detail in future work.

¹our source code: github.com/MarianaGuez/Audio/tree/main/Paper_Reproduction

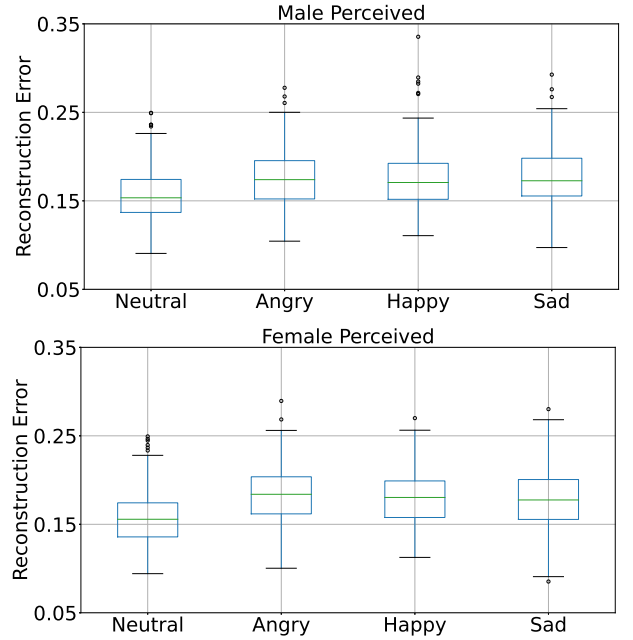


Figure 1: Test-set reconstruction errors by autoencoders trained with neutral utterances, separately shown for men (top) and women (bottom) speech datasets. These results show an x-vector's capability to embed emotion information. The leftmost boxplot shows the reconstruction error of neutral utterances, while the others show the error for angry, happy, and sad utterances. This figure is our replication of a key result by Aldene & Provost (2023) [7].

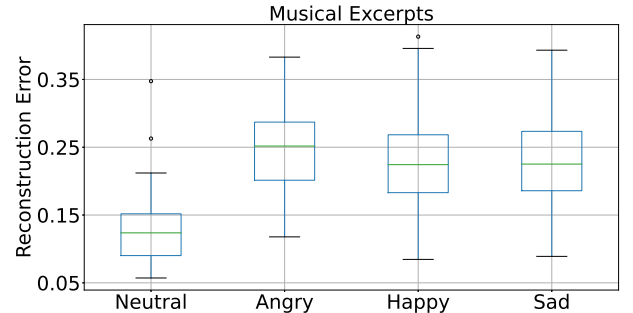


Figure 2: Comparison of the test-set reconstruction error obtained from autoencoders trained on neutral musical excerpts. The first boxplot shows the reconstruction error on neutral music, while the others show the error for angry, happy, and sad musical excerpts. These results average reconstruction across all musical instruments. We observe the same trend reported in [7], but now with music.

4. CONCLUSIONS

Our experiments and preliminary results show that we have been able to replicate the results by Aldene & Provost [7] using their original speech datasets. We also showed that a similar effect is observed if the experiment is carried out using a dataset of musical excerpts we curated. These observations reaffirm that emotional information is embedded in x-vectors across the domains of speech and music. Hence, given that x-vectors in fact contain emotion information, they can be used in future music research targeting tasks such as emotion recognition, or perhaps even generation of emotional and expressive music.

5. ACKNOWLEDGEMENTS

The authors would like to thank the Women in Music Information Retrieval (WiMIR) program for matching Ms. Rodríguez Castañeda and Dr. Roman, allowing them to work together on this project. The presentation of this project at ISMIR 2024 was funded by the ISMIR General Chairs.

6. REFERENCES

- [1] E. Diener, S. Thapa, and L. Tay, “Positive emotions at work,” *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 7, p. 451–477, 2020.
- [2] J. C. Richards, “Exploring emotions in language teaching,” *RELC Journal*, vol. 53, pp. 225 – 239, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:221187211>
- [3] E. Corstorphine, “Cognitive–emotional–behavioural therapy for the eating disorders: working with beliefs about emotions,” *European Eating Disorders Review*, vol. 14, no. 6, pp. 448–461, 2006.
- [4] A. De and S. Mishra, *Augmented Intelligence in Mental Health Care: Sentiment Analysis and Emotion Detection with Health Care Perspective*. Singapore: Springer Nature Singapore, 2022, pp. 205–235. [Online]. Available: https://doi.org/10.1007/978-981-19-1076-0_12
- [5] S. Jain, A. Jain, and M. Jangid, “Review of meta-heuristic techniques for feature selection,” in *Soft Computing: Theories and Applications*, R. Kumar, A. K. Verma, T. K. Sharma, O. P. Verma, and S. Sharma, Eds. Singapore: Springer Nature Singapore, 2023, pp. 397–410.
- [6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [7] Z. Aldeneh and E. M. Provost, “You’re not you when you’re angry: Robust emotion features emerge by recognizing speakers,” *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1351–1362, 2023.
- [8] J. Sager, R. Shankar, J. Reinhold, and A. Venkataraman, “Vesús: A crowd-annotated database to study emotion production and perception in spoken english,” in *Interspeech 2019*, 2019, pp. 316–320.
- [9] F. Font, G. Roma, and X. Serra, “Freesound technical demo,” in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM ’13. New York, NY, USA: Association for Computing Machinery, 2013, p. 411–412. [Online]. Available: <https://doi.org/10.1145/2502081.2502245>
- [10] J. Abeßer, “Idmt-smt-bass-single-track dataset,” Jan. 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.7544099>
- [11] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, “The mtg-jamendo dataset for automatic music tagging,” in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. [Online]. Available: <http://hdl.handle.net/10230/42015>
- [12] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” *Computer Science and Language*, 2019.
- [13] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *INTERSPEECH*, 2018.
- [14] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, “Speech-Brain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.
- [15] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, “Spoken language recognition using x-vectors,” in *The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 105–111.
- [16] J. Williams and S. King, “Disentangling style factors from speaker representations,” in *Interspeech 2019*, 2019, pp. 3945–3949.
- [17] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, “Probing the information encoded in x-vectors,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 726–733.
- [18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.