45. EVE: Emotional Voice Expressions, an acted audiovisual corpus

Elodie Etienne[a], Angélique Remacle[b], Anne-Lise Leclercq[b] and Michaël Schyns[a]

[a] QuantOM, HEC Liège, University of Liège, Liège, Belgium
[b] Research Unit for a life-Course perspective on Health and Education (RUCHE), University of Liège, Liège, Belgium

Type of manuscript: Extended abstract

This work is based on a doctoral thesis in progress

Keywords: speech database; emotions; audiovisual; perceptive study

Extended abstract

In the dynamic interface between humans and technology, the production and perception of emotions are essential for fostering effective communication. The capacity of Artificial Intelligence (AI) systems to accurately perceive, interpret, and react to human emotions is critical, especially in immersive environments with virtual agents capable of detecting emotions. This paper presents the EVE (Emotional Validated Expression) corpus, an acted audiovisual corpus in both English and French currently validated by a perceptive study involving 2,000 listeners. It meticulously embodies the six basic emotions classified by Ekman in 1999, also known as fear, anger, happiness, sadness, disgust, and surprise. It also encompasses four complex emotions: self-confidence, confusion, contempt, and empathy (Perry et al., 2011; Hess et al., 2003; Hareli et al., 2018; Geer et al.,2000). The addition of the four complex emotions was guided by the Warmth and Competence model (Fiske et al., 2007), a well-established framework that assesses how individuals perceive others using the dimensions of Warmth (friendliness and trustworthiness) and Competence (efficiency and skill). With its free accessibility under an open license, the EVE corpus aims to become an indispensable asset for research and innovation in AI, Robotics, and Smart Interfaces, thereby contributing to the evolving narrative of Industry 4.0.

The development of the EVE corpus was driven because many existing emotional speech corpora used for Speech Emotion Recognition (SER), such as CREMA-D(Cao et al., 2014), IEMOCAP (Busso, 2008), and RAVDESS (Livingstone, 2018) in English, EmoV-DB (Adigwe et al., 2018) and EmoVox (Schrerer, 2013) available in both English and French, along with CaFE (Gournay et al.) and Oréau (Kerkeni et al., 2020) in French, often exhibit shortcomings including limited data volume, restricted emotional variety, narrow actor diversity, absence of phonetic balance, and sometimes lack of validation (see Table 1). Among other available datasets, the HUME dataset (Cowen et al., 2019) offers an extensive collection of over 40,000 samples. However, it might remain inaccessible for many researchers due to its cost or specific usage restrictions.

The primary goal of the EVE corpus is to fill the gap in the availability of validated and high-quality SER databases for both English and French. To achieve this, the corpus provides a diverse collection of phonetically balanced sentences produced in diverse emotions, ensuring equal representation of genders among the actors. Moreover, the corpus is undergoing validation through a comprehensive perceptive study that engaged 2,000 listeners to assess the emotion conveyed in each recording.

The EVE corpus was developed by emphasising the production of high-quality recordings. The recording sessions took place in a professional soundproof room. Actors were equipped with a high-quality microphone headset connected to an external sound card. A tracking camera was positioned to record the actors' facial expressions and upper body movements. For each language, ten actors performed each sentence from the Harvard (Rothauser et al., 1969) and FHarvard (Aubanel et al., 2020) lists, known for their phonetic balance, with every emotion with two different intensities (low and high) and a neutral state. To maintain consistency, each actor repeated each sentence an emotion combination twice (repetition technique in acting). This comprehensive approach resulted in 4,100 high-definition audiovisual recordings per language, ranging from 2 to 10 seconds.

For validation, the EVE corpus is undergoing a detailed perceptual study in each language to assess the emotions perceived in 2,000 selected recordings (i.e., the second attempt of each emotional recording). This involves 1,000 speakers of English and French respectively, (covering, in each language, a broad linguistic and cultural diversity), evaluating the recordings via an online platform. The study presents listeners with a randomised selection of recordings, asking them to identify and rate their confidence in the emotions depicted, first using audio cues only and then the audiovisual content. Each listener evaluates 50 clips randomly, ensuring comprehensive corpus coverage and adequate listener responses for statistical reliability.

The foundational hypotheses were formulated based on a comprehensive review of the current literature and initial investigations into the expression and recognition of emotions. Research indicates that basic emotions, as defined by Ekman (1999), are generally recognised universally across various cultures and languages, as evidenced by studies conducted by Monroy et al. (2022) and Cowen et al. (2019). This recognition extends to computational models. However, recognising complex emotions presents a more significant challenge due to their subtler manifestations and heavier reliance on context, as Cowen et al. (2019) discussed. Recent research by Tomar (2024) reaffirmed the significant role of visual cues in enhancing emotion recognition, offering vital supplementary information to auditory signals. This leads to the following hypotheses:

- H1: Basic emotions will have higher recognition (H1a) and higher (H1b) confidence rates in both languages than complex emotions.
- H2: Visual cues will significantly enhance the recognition (H2a) and confidence (H2b) rates, providing additional information to auditory cues.
- H3: Emotions expressed at higher intensities will be recognised more easily, as more pronounced emotional expressions tend to be clearer and more discernible.

The study is anticipated to conclude by early June, when a comprehensive assessment of the corpus and its applicability will be available. The analysis will delve into the recognition rates of emotions, examining these rates regarding the complexity of emotion, the language (French vs. English), and the presentation modality (audio vs. audiovisual). Preliminary findings are already aligned with our initial hypotheses.

The EVE corpus represents a pivotal advancement in SER, combining technological innovation with a focus on human-centric communication, which aligns with Industry 4.0's vision of integrating intelligent systems into everyday human activities. Through the meticulous development of the corpus and its extensive perceptive validation, the project is set to enhance AI's emotional intelligence, promising to transform human-computer interactions across various sectors such as healthcare, education, and entertainment using virtual environments and virtual agents capable of detecting the emotions of the immersed person. This endeavour aims to make AI interactions more

natural and intuitive and underscores the corpus's potential to become a critical resource for the research and development community.

Table 1. Most popular open-access corpora

Name	Modality	Type	Nb. speakers	Nb. emotions	Nb. intensities	Phonetical balance	Perceptual study
French							
CaFE (Gournay *et al.* , 2018)	Audio	Acted	12	6+neutral	1	x	x
EmoV-DB (Adigwe *et al.,2018)*	Audio	Acted	1	4+neutral	1	v	v
EMOVOX (Scherer, 2013)	Audio	Induced Acted	54	2	1	x	x
Oréau (Kerkeni *et al.*, 2020)	Audio	Acted	32	7	1	x	v
English							
CREMA-D (Cao *et al.*, 2014)	Audio-visual	Acted	1	5+neutral	3	x	v
EmoV-DB (Adigwe *et al.,2018)*	Audio	Acted	4	4+neutral	1	v	v
EMOVOX (Scherer, 2013)	Audio	Induced Acted	16	2	1	x	x
IEMOCAP (Busso, 2008)	Audio-visual	Acted	10	8+neutral	1	x	x
RAVDESS (Livingstone, 2018)	Audio-visual	Acted	24	7+neutral	2	x	v

Acknowledgements
The authors gratefully thank the reviewers for their insightful comments and suggestions, which improved the quality of this work.

References

Aubanel, V., Bayard, C., Strauss, A., & Schwartz, J. L. (2020). The Fharvard corpus: A phonemically-balanced French sentence resource for audiology and intelligibility research. Speech Communication, 124, 68-74.

Adigwe, A., Tits, N., Haddad, K. E., Ostadabbas, S., & Dutoit, T. (2018). The emotional voices database: Towards controlling the emotion dimension in voice generation systems. arXiv preprint arXiv:1806.09514.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S.(2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42, 335-359.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014).Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), 377-390.

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. Nature human behaviour, 3(4), 369-382.

Ekman, P. (1999). Basic emotions. Handbook of cognition and emotion, 98(45-60), 16.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. Trends in cognitive sciences, 11(2), 77-83.

Geer, J. H., Estupinan, L. A., & Manguno-Mire, G. M. (2000). Empathy, social skills, and other relevant cognitive processes in rapists and child molesters. Aggression and violent behavior, 5(1), 99-126.

Gournay, P., Lahaie, O., & Lefebvre, R. (2018, June). A canadian french emotional speech dataset. In Proceedings of the 9th ACM multimedia systems conference (pp. 399-402).

Hareli, S., Halhal, M., & Hess, U. (2018). Dyadic dynamics: The impact of emotional responses to facial expressions on the perception of power. Frontiers in psychology, 9, 364852.

Hess, U. (2003). Now you see it, now you don't--the confusing case of confusion as anemotion: Commentary on Rozin and Cohen (2003).

Kerkeni, ML. Cledern, C., Serrestou, Y., & Raood, Y. (2020). French emotional speech database-oréau.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.

Monroy, M., Cowen, A. S., & Keltner, D. (2022). Intersectionality in emotion signaling and recognition: The influence of gender, ethnicity, and social class.Emotion, 22(8), 1980.

Perry, P. (2011, October). Concept analysis: Confidence/self confidence. In Nursing forum-(Vol. 46, No. 4, pp. 218-230). Malden, USA: Blackwell Publishing InC.

Rothauser, E. H. (1969). IEEE recommended practice for speech quality measurements. IEEE Transactions on Audio and Electroacoustics, 17(3), 225-246.Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. Computer Speech & Language, 27(1), 40-58.

Tomar, P. S., Mathur, K., & Suman, U. (2024). Fusing facial and speech cues for enhanced multimodal emotion recognition. International Journal of Information Technology, 1-9