EVE: Emotional Validated Expressions, an acted audiovisual corpus

Elodie Etienne
HEC, ULiège
Liège, Belgium
elodie.etienne@uliege.be

Angélique Remacle
Speech-Language Pathology Department, ULiège
Liège, Belgium

Anne-Lise Leclercq
Speech-Language Pathology Department, ULiège
Liège, Belgium

Michaël Schyns
HEC, ULiège
Liège, Belgium

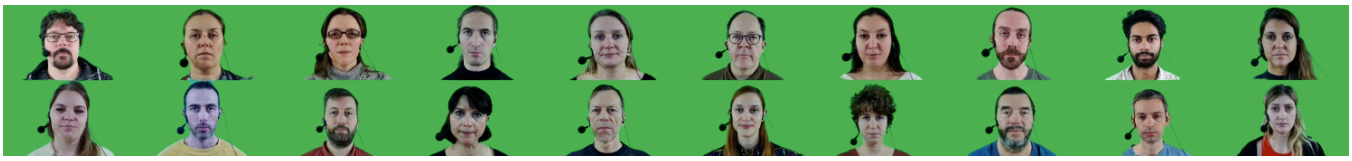


Figure 1: Actors of the corpus

ABSTRACT

This paper presents the creation and perceptual validation of the EVE corpus, a resource, in English and French, for speech emotion recognition of audio and audiovisual content and for the generation of verbal and non-verbal behaviour. For each language, ten actors performed 10 linguistically and semantically neutral sentences with different emotions (fear, anger, happiness, sadness, disgust, surprise, confidence, confusion, contempt, empathy) and a neutral condition. Each was expressed at two arousal levels, with two trials per level. The emotional content of the corpus was perceptually validated by 600 participants per language. The corpus is accessible through this link: https://doi.org/10.58119/ULG/VREIOB.

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; • **Information systems** → *Multimedia information systems*; • **Human-centered computing** → **Interaction paradigms**.

KEYWORDS

emotions, multimodal, speech, audio, audiovisual, data, corpus, perceptual study

ACM Reference Format:
Elodie Etienne, Angélique Remacle, Anne-Lise Leclercq, and Michaël Schyns. 2025. EVE: Emotional Validated Expressions, an acted audiovisual corpus. In *ACM International Conference on Intelligent Virtual Agents (IVA'25), September 16–19, 2025, Berlin, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3717511.3749303

IVA '25, September 16–19, 2025, Berlin, Germany

ACM ISBN 979-8-4007-1508-2/2025/09
https://doi.org/10.1145/3717511.3749303

1 INTRODUCTION

Intelligent virtual agents (IVAs) are increasingly integrated into interactive systems designed to engage users through natural communication. To participate in meaningful human-agent interactions, IVAs must be capable not only of interpreting the user's verbal and non-verbal signals but also of producing expressive behaviours that reflect appropriate emotional states. Enabling IVAs to detect and generate emotional speech or facial expressions enhances their social presence, fosters user trust, and improves the overall interaction experience [10]. This capability is increasingly achieved through artificial intelligence, which requires large, high-quality datasets for training and evaluation.

In human communication, emotions emerge not only through basic states such as fear, anger, happiness, sadness, disgust, and surprise [9, 27], but also through more complex affective states. Emotions like self-confidence [20], confusion [15], contempt [14], and empathy [12] are often essential for nuanced communication. These emotions are particularly relevant in domains where IVAs are expected to interpret subtle cues and respond appropriately [10].

Achieving this goal requires access to annotated emotional speech data that is both diverse and reliable. Such data is essential not only for rule-based approaches but also for data-driven methods, where models learn to recognise and generate expressive behaviours based on annotated emotional corpora. In particular, emotional corpora for IVA must support two core needs: recognising the emotional states of users from their speech and synthesising emotionally appropriate responses for the agent. The creation of the EVE (Emotional Validated Expressions) corpus was motivated by the scarcity or limited accessibility of publicly available high-quality databases in both French and English. Indeed, many existing speech audiovisual corpora are limited in scope or prohibitively expensive. For instance, the Hume AI database [7], despite its extensive data volume (400,000 recordings), may be inaccessible to many researchers due to its cost or usage conditions. Additionally, its restriction to

only five different sentences results in an unbalanced distribution of expressions. Similarly, databases commonly used in SER research in English, often do not simultaneously fulfil various crucial criteria such as emotional diversity [24], different levels of arousal [4, 16, 21, 24], phonetic balance [4, 5, 18, 21, 24], sufficient amount of speakers [16, 21] and perceptual validation of intended emotions by a statistically significant number of participants [4, 21, 24]. The challenge intensifies for French databases which face similar issues in terms of emotional diversity[24], speakers diversity [17, 24], phonetic balance [13, 17], and validity through perceptual studies [2, 13] if not due to recording quality (e.g., control of microphone types, background noise, or studio vs. in situ recordings). All the referenced databases present various limitations, highlighting the need for datasets that address these shortcomings to enable more effective recognition and generation.

2 THE EVE CORPUS

2.1 Corpus Creation: Data Collection

This study was approved by the Ethics Committee of the Faculty of Psychology, Speech Therapy, and Educational Sciences of the University of Liège (file number: 2223-087).

Beyond the neutral condition, the emotions selected for the corpus were chosen to encompass a broad spectrum, including six basic emotions and four complex emotions, ensuring both foundational and nuanced emotional states are represented for diverse research applications (see section 1). Each emotion was expressed at two levels of arousal (low and high). For example, for sadness, low arousal was like having a lump in the throat, while strong arousal resembled almost bursting into tears [13].

The sentences of this corpus were carefully chosen to ensure neutral linguistic and semantic content, avoiding any emotional connotations. From a phonemic perspective, the goal was to achieve phonetic completeness and balance. To this end, a phonetically balanced list of sentences was selected from established corpora: the first ten sentences of the Harvard Sentences for English [23] and the FHarvard corpus for French [1].

The EVE corpus was created using ten actors, native or near-native speakers, consisting, for each language, of five males and five females (see Figure 1), all recruited voluntarily and possessing expertise in the dramatic arts for more than three years.

The recording sessions were conducted individually in a professional soundproof room to ensure optimal audio quality. The actors stood approximately 20 cm away from a green wall. They wore a headset microphone (AKG C 54) positioned around 5 cm from their mouths and connected to a microphone preamp (Focusrite iTrack Solo). This setup was linked to an Apple MacBook Pro laptop (2.3 GHz Intel Core i5 Dual CoreN) running Camtasia software (version 22.5.4) to enable recording at high definition (1280x720). In addition, a tracking camera (Obsbot Tiny OWB-2004-CE) was placed in front of them to capture their faces and upper torsos. Each recording session was structured to last two hours per actor. The actors were instructed to perform each sentence with every emotion at both arousal levels, with two trials for each level, incorporating techniques from the Meisner acting method [6] to enhance emotional authenticity and spontaneity. Indeed, it was observed that emotions tended to be more accurately conveyed on the second trial, possibly due to the emphasis on repetition and response [25]. A random sequence of sentences and emotions was assigned for each actor to prevent any order effects. Rigorous quality control measures were implemented, focusing on correcting mispronunciations, reducing hesitations and background noise, and ensuring a consistent recording environment.

In total, for each language, 4,100 utterances were recorded: 100 neutral utterances (10 actors × 10 sentences), and 4000 emotional utterances (10 actors × 10 sentences × 10 emotions × 2 arousal levels × 2 trials). For each sentence in the corpus, two files were exported: one stereo video containing both audio and visual cues (format H.264, image resolution of 1,920 x 1,080, 16:9 aspect ratio, 60 fps, with a .MP4 extension), and one mono audio file (format Waveform Audio, 16-bit, 44.1 kHz, with a .WAV extension).

The EVE corpus comprises 8,200 high-quality recordings, evenly split between English and French. The English corpus totals 3 hours, 46 minutes, and 50 seconds, with individual file durations ranging from 2 to 8.12 seconds. The French corpus totals 4 hours, 3 minutes, and 45 seconds, with durations from 2.06 to 11.6 seconds.

2.2 Corpus Validation: Perceptual Study

An online perceptual experiment was conducted on the corpus to assess the presence of the portrayed emotions in the audio recordings. For the perceptual study, 2000 recordings were selected for each language. They correspond to the second trial of each utterance. The goal was to assess the perceived emotions using first the audio and then the audiovisual stimuli. Thus, for each audio, participants were asked to identify the emotion being depicted and rate their confidence in their identification. Additionally, they were asked whether they would have preferred to select "no idea" instead of choosing an emotion and whether they hesitated between several possible emotions. In the latter case, they were prompted to specify the emotions they considered, ranking them from the most to the least probable. This process was then immediately repeated for the corresponding audiovisual recording, with participants' previous choices from the audio-only evaluation preselected to help them recall their initial responses. However, they were free to modify their selections if the addition of visual information altered their perception of the emotion or confidence level.

Participants for the perceptual study were recruited through Prolific. Participants were compensated £6 for an estimated 1-hour study, although the median completion time for this study was 40 minutes. To ensure data quality, participants were rigorously screened using platform-based filters based on their commitment to previous studies. Random responses were filtered out through automated and manual validation checks. The platform implemented controls, with hidden mechanisms flagging inattentive behaviour, ensuring high-quality annotations throughout.

The study involved 1,200 participants, evenly divided between those fluent in English and French, who were screened for language proficiency (using filters available in Prolific). Participants were also required to reside in countries where the respective language is an official language. Ages ranged from 18 to 78 years for the English group (median age: 34), comprising 55.5% females, 43.3% males, and 1.2% who did not specify their gender. For the French

group, ages ranged from 18 to 75 years (median age: 29), with 49% females, 50% males, and 1% preferring not to disclose their gender. The comprehensive online perceptual experiment results are provided in CSV format, where each row corresponds to an evaluation by a participant for a specific file. Another CSV file summarizes perceived emotions for each recording, with rows representing individual recordings and columns indicating the proportions of perceived emotions. CSV files specific to each parameter are also provided, ensuring a detailed analysis and offering all the necessary information of this corpus.

3 RESULTS OF THE PERCEPTUAL STUDY

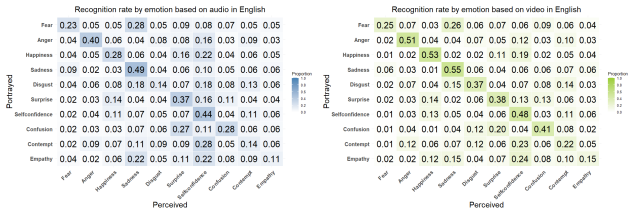
In terms of recognition rates per emotion, Figures 2 and 3 show that visual cues significantly enhance emotion recognition rates in both English and French. Even with audio-only data, all emotions are recognized above the random recognition rate (i.e., 0.1), with some emotions being more accurately identified than others. Specifically, sadness, anger, self-confidence, and surprise tend to be recognized more reliably in English, while in French, sadness, anger, surprise, self-confidence, and confusion show higher recognition rates compared to other emotions.

For confidence levels per modality, results highlight a significant improvement brought by visual cues (Paired t-tests confirm the statistical difference for both English, ($T = -65.64, df = 29999, p < 0.001$) and French ($T = -74.96, df = 29999, p < 0.001$)), suggesting that visual information not only boosts recognition rates but also increases confidence in emotion identification. When analysing emotions individually, all tested emotions exhibit significant differences. Disgust and happiness exhibit the largest confidence gains in both languages. In English, confusion, anger, and self-confidence follow, while in French, contempt, confusion, and self-confidence follow.

When examining recognition rates across arousal levels for all emotions combined, there is a trend where emotions expressed with higher arousal are recognized more accurately. However, even at lower arousal levels, recognition rates remain consistently above the random choice threshold, indicating that emotions can still be reliably identified regardless of the arousal level. Paired t-tests suggest a tendency for improved recognition at higher arousal levels, though the statistical difference is not strongly significant for either English ($T = -1.9762, df = 9, p = .07955$) or French ($T = -2.1722, df = 9, p = 0.05791$). Chi-squared tests comparing recognition rates for each emotion individually further indicate that, in English, only sadness, self-confidence, and empathy do not show a significant difference (for a significance of .05) in recognition rates between arousal levels. In French, with the same level of significance, only confusion and empathy are non-significant, while all other emotions exhibit a significant effect of arousal levels.

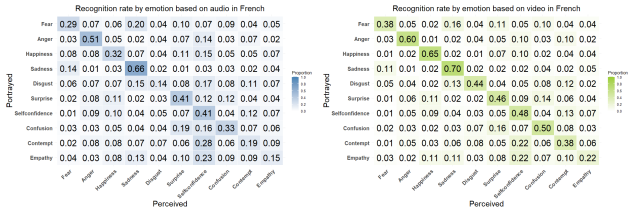
The ranking of recognition rates per actor differs between audio-only and audiovisual conditions, highlighting individual differences in emotional expression across modalities.

Lastly, for each language, the Marascuilo Procedure was applied to test for statistical differences between the recognition rates by sentence. As expected, no difference was found, as the sentences were designed to be emotionally neutral. Yet, this is an important result that validates our methodological procedure.



(a) Audio **(b) Audiovisual**

Figure 2: Confusion matrices of recognition rate based on audio (a) and audiovisual (b) data in English.



(a) Audio **(b) Audiovisual**

Figure 3: Confusion matrices of recognition rate based on audio (a) and audiovisual (b) data in French.

4 DISCUSSION

This paper presents the creation and validation of the EVE corpus, focusing on emotional audio and audiovisual speech in English and in French. Built from phonemically balanced sentences, the corpus enables cross-linguistic and multimodal analyses. A key strength lies in its perceptual validation by a large, representative sample. This revealed consistent trends: some emotions are better recognised than others, some are frequently confused, and probabilistic emotion labels can be assigned to each recording. Similarities across languages were observed. EVE also addresses several gaps in existing corpora, such as emotional diversity, phonetic balance, and perceptual validation, while remaining usable for both recognition and synthesis tasks—especially in ML pipelines.

Some limitations remain. First, the dataset is relatively small, though high-quality small datasets can outperform larger ones in deep learning [22]. Second, the focus on English and French limits broader cultural coverage. While the use of basic emotions ensures consistency with existing emotional databases, one can argue that those are not universal [3]. Similarly, the use of a single acted sentence may lack the richness of context and spontaneous speech [11, 19]. Finally, variability in emotion perception across language proficiencies and cultural backgrounds remains a challenge [8, 26]. Further perceptual studies could guide the design of more adaptive IVAs.

Despite these constraints, EVE already provides valuable, generalisable insights and a solid foundation for future research and emotion-aware IVA development.

REFERENCES

[1] Vincent Aubanel, Clémence Bayard, Antje Strauss, and J-L Schwartz. 2020. The Fharvard corpus: A phonemically-balanced French sentence resource for audiology and intelligibility research. *Speech Communication* 124 (2020), 68–74.

[2] Tanja Bänziger, Hannes Pirker, and K Scherer. 2006. GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. In *Proceedings of LREC*, Vol. 6. 15–019.

[3] Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain*. Pan Macmillan.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. [n. d.]. IEMOCAP: interactive emotional dyadic motion capture database. 42, 4 ([n. d.]), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[5] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.

[6] Drama Classes and Performing Arts School. 2025. Meisner Technique. https://www.dramaclasses.biz/meisner-technique Accessed: 31-Jan-2025.

[7] Alan S Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour* 3, 4 (2019), 369–382.

[8] David Efron. 1972. *Gesture, Race and Culture: A Tentative Study of the Spatio-temporal and "linguistic" Aspects of the Gestural Behavior of Eastern Jews and Southern Italians in New York City, Living Under Similar as Well as Different Environmental Conditions*. Mouton. https://books.google.be/books?id=yMY6uwEACAAJ

[9] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. 2013. *Emotion in the human face: Guidelines for research and an integration of findings*. Vol. 11. Elsevier.

[10] Elodie Etienne, Marion Ristorcelli, Sarah Saufnay, Aurélien Quilez, Rémy Casanova, Michael Schyns, and Magalie Ochs. 2024. A Systematic Review on the Socio-affective Perception of IVAs' Multi-modal behaviour. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents* (GLASGOW, United Kingdom) *(IVA '24)*. Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. https://doi.org/10.1145/3652988.3673943

[11] Steffi Frigo. 2006. The relationship between acted and naturalistic emotional corpora. In *Workshop" Corpora for research on emotion and affect". 5th International Conference on Language Resources and Evaluation (LREC'2006)*. 34–36.

[12] James H Geer, Laura A Estupinan, and Gina M Manguno-Mire. 2000. Empathy, social skills, and other relevant cognitive processes in rapists and child molesters. *Aggression and violent behavior* 5, 1 (2000), 99–126.

[13] Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 2018. A canadian french emotional speech dataset. *Proceedings of the 9th ACM Multimedia Systems Conference* (2018). https://api.semanticscholar.org/CorpusID:49644035

[14] Shlomo Hareli, Mano Halhal, and Ursula Hess. 2018. Dyadic dynamics: The impact of emotional responses to facial expressions on the perception of power. *Frontiers in psychology* 9 (2018), 1993.

[15] Ursula Hess. 2003. Now you see it, now you don't–the confusing case of confusion as an emotion: Commentary on Rozin and Cohen (2003). (2003).

[16] Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK* (2014).

[17] Leila Kerkeni, Catherine Cleder, Ser-Restou Youssef, and Kosai Raoof. 2020. French emotional speech database-oréau.

[18] Steven R. Livingstone and Frank A. Russo. [n. d.]. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. 13, 5 ([n. d.]), e0196391–e0196391. https://doi.org/10.1371/journal.pone.0196391

[19] Shae D Morgan and Bailey LaPaugh. 2025. Methodological Stimulus Considerations for Auditory Emotion Recognition Test Design. *Journal of Speech, Language, and Hearing Research* (2025), 1–16.

[20] Patricia Perry. 2011. Concept analysis: Confidence/self-confidence. In *Nursing forum*, Vol. 46. Wiley Online Library, 218–230.

[21] M. Kathleen Pichora-Fuller and Kate Dupuis. 2020. Toronto emotional speech set (TESS). https://doi.org/10.5683/SP2/E8H2MF

[22] Ishfaq Hussain Rather, Sushil Kumar, and Amir H Gandomi. 2024. Breaking the data barrier: a review of deep learning techniques for democratizing AI with small datasets. *Artificial Intelligence Review* 57, 9 (2024), 226.

[23] EH Rothauser. 1969. IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics* 17, 3 (1969), 225–246.

[24] Klaus R Scherer. 2013. Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language* 27, 1 (2013), 40–58.

[25] Sheldon Schiffer. 2019. How actors can animate game characters: integrating performance theory in the emotion model of a game character. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 15. 227–229.

[26] Stella Ting-Toomey and Tenzin Dorjee. 2018. *Communicating across cultures*. Guilford Publications.

[27] University of West Alabama Online. 2019. Our Basic Emotions Infographic. https://online.uwa.edu/infographics/basic-emotions/. Accessed: 2025-06-09.