Speech Emotion Recognition for Public Speaking Training in Virtual Reality

1st Saufnay Sarah
HEC Liège
University of Liège
Liège, Belgium
sarah.saufnay@uliege.be

2nd Etienne Elodie
ALMAnaCH
INRIA
Paris, France
elodie.etienne@inria.fr

3nd Vanmechelen Thibaud
HEC Liège
University of Liège
Liège, Belgium

4nd Schyns Michaël
HEC Liège
University of Liège
Liège, Belgium
m.schyns@uliege.be

***Abstract*—Public speaking often triggers strong emotional responses that influence both the speaker's performance and the audience's reactions. By immersing users in realistic speaking contexts in front of a virtual audience, Virtual Reality (VR) constitutes an effective training solution for such skills. To enable the development of responsive virtual agents, this work-in-progress presents research aimed at designing a Speech Emotion Recognition (SER) system compatible with VR environments. The prediction of ten emotions is targeted using various machine learning approaches, relying solely on the speaker's speech signal. To train the models, a bilingual acted corpus was used. Thanks to a perceptual validation study, the corpus was annotated with the emotions human raters effectively perceived. The proposed methodology is presented along with preliminary results.**

***Index Terms*—speech emotion recognition, emotions, affective computing, virtual reality, public speaking**

I. Introduction

Speaking in front of an audience, whatever its size or the associated context, can be stressful for many people. Such situations often trigger strong emotional reactions [13], with speakers experiencing fluctuations between anxiety and self-confidence [18]. These emotions evolve dynamically throughout the speech, as they are strongly influenced by the audience's reactions [19] [26]. The ability to regulate one's feelings is essential, not only for traditional public speaking tasks but also for situations that require persuasion, empathy, or support towards others. In professional interactions, expressing positive emotions such as happiness or confidence can enhance persuasiveness and credibility, and can even become a strategic argument [28] [33]. In the healthcare sector, showing empathy and maintaining supportive interactions with patients can even contribute to the recovery process [14] [31]. Conversely, feelings such as anger or fear may be deliberately masked to preserve positive impressions, thereby fostering workers' well-being by helping them avoid conflicts [20]. During crisis communication, emotional regulation is crucial for strengthening message clarity [7] and reducing adverse effects [32]. In all these contexts, emotional regulation shapes how the speaker is perceived, thereby strongly influencing social relationships.

Thanks to its immersive capabilities, Virtual Reality (VR) offers a particularly promising solution for training these skills [3] [27]. Indeed, VR enables users to immerse themselves in 3D environments that replicate various public speaking contexts, allowing them to practice in front of a virtual audience [21] [24] [26]. However, to realistically replicate real-life presentations and improve training effectiveness, systems should integrate a responsive audience that behaves realistically throughout the presentation [6]. This requires real-time detection of the speaker's emotional state and performance, which determines the audience reactions [9]. While emotions are expressed through multiple modalities, including facial expressions [17], physiological responses [25], linguistic content, and paralinguistic cues [15] [23], most of these indicators cannot be tracked with a standard VR headset. In fact, few VR devices integrate facial tracking, and external sensors are required for physiological monitoring. In any case, the trackers are often not sufficiently precise for accurate emotional analysis. In contrast, the user's speech remains accessible during VR immersion and constitutes a powerful and reliable source of information for emotion detection [2] [11].

The present work addresses this objective by developing a Speech Emotion Recognition (SER) system designed for VR public speaking training that relies exclusively on acoustic cues. By focusing on prosodic and spectral properties rather than lexical content, this approach aims to achieve generalization across public speaking contexts. The following sections present the proposed methodology and preliminary results. The emotional corpus used to train the machine learning algorithms is also introduced, and an analysis of the extracted acoustic features is conducted to guide feature selection.

II. Speech Emotion Recognition

Multiple frameworks have been proposed to describe human emotions [2]. Some define them as discrete categories, such as fear, anger, sadness, happiness, disgust, and surprise [8] [29]. In contrast, others describe emotions as continuous variations within a multidimensional space [12], typically along valence, arousal, and dominance axes [30]. Both approaches have guided the development of SER models.

In speech, emotions are expressed through multiple cues, including linguistic, contextual, and acoustic ones. Among these, prosodic features, such as pitch, intensity, speech rate, and voice quality, play a key role in conveying emotions [11]. For instance, higher pitch and energy are typically associated with high-arousal emotions such as anger or happiness, whereas

2771-7453/26/$31.00 ©2026 IEEE
DOI 10.1109/AIxVR67263.2026.00028

lower pitch and intensity are characteristic of fear or sadness [4]. These paralinguistic cues provide a solid basis for emotion detection, making them even more relevant when other modalities (*e.g.* nonverbal and physiological) are unavailable or of insufficient quality, as is the case in most VR headsets. Moreover, the extraction and analysis of such features have been facilitated by dedicated toolkits, like openSMILE [11], which enable reliable analysis of voice signals.

The extraction of acoustic features has supported a wide range of SER systems [2]. Approaches based on Support Vector Machines (SVMs), k-Nearest Neighbors (KNNs), or Random Forests (RFs) demonstrated that acoustic features could reliably predict emotions [2]. More recently, deep learning architectures and transformer-based representations, such as HuBERT [1], have further advanced SER performance [5]. Despite these improvements, several challenges remain [2]. Emotion expression and perception indeed vary between individuals [10], cultures [16], and languages [22], complicating the generalization. Moreover, faster and more effective SER systems are needed, primarily by using more accurate data [2]. An outstanding issue is that most existing SER systems are trained on acted emotions, which may not reflect how people effectively perceive them. However, in communicative contexts, such as public speaking, vocal cues may be interpreted differently by different listeners [10]. To address these limitations, the present work proposes a perceptually oriented approach in which models are trained not to recognize acted emotions but emotions as perceived by human raters, thereby supporting realistic audience design in VR.

III. EVE CORPUS

Developing SER systems requires access to high-quality annotated emotional corpora for model training. Due to the lack of freely available databases with the properties needed for this project, the Emotional Validated Expressions (EVE) corpus has been created [10]. EVE is an audiovisual, bilingual acted corpus comprising 8,200 recordings, equally distributed between French and English. For each language, ten actors expressed ten distinct emotions through ten phonetically balanced sentences. This database covers a wide range of affective states, including the six basic Ekman emotions (*i.e.*, anger, disgust, fear, happiness, sadness, and surprise) [8], as well as four additional ones particularly relevant to public speaking contexts (*i.e.*, confusion, contempt, empathy, and self-confidence) [10]. Although highly relevant, these emotions constitute more complex affective states than those described by Ekman. Each actor had two trials to portray these emotions, at two distinct arousal levels (*i.e.,* low and high), yielding 4,000 emotional recordings per language. A neutral condition was also recorded for each actor, adding 100 (*i.e.*, 10 actors x 10 sentences) recordings per language. Finally, the dataset comprises 3 h 46 min 50 s of English recordings and 4 h 03 min 45 s of French recordings, with file durations ranging from 2 to 8.12 s and from 2.06 to 11.6 s, respectively [10].

As presented in [10], the EVE corpus stands out from existing emotional databases for its inclusion of French recordings, its diversity of emotions, and, most importantly, its validation through a perceptual study. In fact, a large-scale study was carried out involving 1,200 participants, who evaluated 2,000 audio recordings per language (*i.e.,* the second trial for each emotion). Participants were asked to (1) identify the perceived emotion for each recording, (2) rate their confidence, and (3) optionally indicate emotions they hesitated between. This perceptual validation study provides valuable insight into how affective states are actually perceived by human audiences, rather than how actors intended them. This is particularly relevant as perception matters more than the speaker's intentions during public speaking tasks. The present work then relies on this emotional corpus.

IV. METHODOLOGY

Given the limitations of VR headset tracking, the SER system will focus exclusively on acoustic features. Accordingly, only the audio recordings from the EVE corpus and their associated perceptual evaluations will be used. Two separate models will first be developed for French and English, although a multilingual approach is considered for future work.

A. Features extraction and analysis

Acoustic features were extracted using the GeMAPS configuration proposed by openSMILE [11]. This set of 62 prosodic and spectral features includes frequency-related parameters (*i.e.*, Pitch, Jitter, Formant frequencies), energy-related parameters (*i.e.*, Shimmer, Loudness, Harmonics-to-Noise Ratio (HNR)), spectral measures (*i.e.*, Alpha Ratio, Hammarberg Index, Formant relative energy, Harmonic differences), and temporal features (*i.e.*, loudness peak rate, voiced and unvoiced regions information). From these acoustic measures, a range of derived functionals (*e.g.* mean, standard deviation, and percentiles) were computed to provide further information. This complete set of parameters can then be used to represent the prosodic energy, voice quality, and spectral structure associated with emotional expressions. The extraction process was applied to the 2,000 evaluated audio recordings for each language in the EVE database. To handle redundancy among the extracted variables, correlation matrices and Principal Component Analysis (PCA) were used to evaluate feature interdependence and structure (see Section V). These analyses will guide the selection of a reduced and interpretable subset of features. Through this dimensionality reduction, the aim is to improve the model's efficiency while minimizing the risk of overfitting [2], especially given the moderate dataset size (*i.e.*, 2,000 labeled recordings per language). Indeed, as highlighted in [2], carefully selecting the feature set used for SER greatly influences the obtained results.

B. Labeling approaches

Rather than using acted emotions, the present study relies on perceptual *soft labels*, derived from the participants' evaluations of the EVE corpus. As previously mentioned, listeners identified the perceived emotion for each audio recording and indicated their level of confidence in that assessment, thereby

providing the basis for the corpus annotation. In that context, the labeling process requires particular attention due to the uneven distribution of perceived emotions among participants. A first option would be to use all individual responses from the perceptual study, assigning a single perception value to each participant–recording pair. While this increases the amount of data, it also introduces conflicting information, since a single recording may be associated with multiple perceived emotions across participants. Despite the size advantage associated with this approach, it was still excluded from the analysis but will be considered in future work. Instead, to better reflect the variability in human emotion detection, probability vectors were derived from the perceptual data and assigned to each recording, resulting in probability distributions over the ten considered emotions (*i.e.*, soft labels). Each sample thus corresponds to a single audio file from which openSMILE features were extracted, paired with a single soft-label distribution, even though individual participant ratings may differ. Similarly, the confidence scores were used to adjust these distributions, giving more weight to judgments with greater confidence. This strategy was preferred over a single-label approach based on the most recognized emotion. Such a simplification would result in a loss of information about the complexity of human perception and was therefore not adopted. Instead, learning from these distributions enables the model to approximate how listeners actually perceive emotions.

C. Labels description

Three types of soft labels were then considered. The first soft labels type corresponds to *raw-votes*, obtained by normalizing the frequency of each perceived emotion across participants. Each recording is thus represented by a ten-dimensional probability vector that directly reflects the distribution of judgments. A more straightforward approach would have been to assign each audio file to the most commonly perceived emotion (*i.e.*, the distribution's mode). Whereas this single-label approach was excluded from this analysis, the mode can still be extracted from the predicted distribution vector to obtain a single emotion as output. The second label type incorporates *individual confidence weighting*, where each participant's vote is weighted by their self-reported confidence. This approach gives the more confidently selected emotions a greater influence on the final probability distribution, thereby reducing the impact of uncertain responses. The third soft-label approach uses *average-confidence weighting*, where each emotion's weight is the average confidence across all listeners who selected it for that recording. This approach gives greater weight to emotions that are consistently judged across listeners. These approaches will be considered and compared.

V. Features selection

A. Correlation analysis

First, correlation matrices were computed for the extracted openSMILE acoustic features from each recording, both in French and English. This highlighted patterns across languages, with strong dependencies ($|r| > .8$) observed within feature families (*i.e.*, between basic acoustic features and their derived functionals) [11]. Overall, 38 features in English and 42 in French showed at least one strong correlation ($|r| > .8$) with another feature, indicating redundancy within the GeMAPS set. F0-related measures and loudness statistics showed the highest within-family correlations, suggesting that they capture similar information. In contrast, cross-family correlations were generally weak, confirming that distinct families convey complementary information. Only a few strong correlations between families emerged, notably between pitch and voice-quality descriptors (*e.g.* F0 and HNR), as well as between pitch and spectral slope measures (*e.g.* F0 and spectral slope). In summary, this correlation structure supports a reduction strategy that preserves informative features from each family, while minimizing redundancy associated with highly correlated functionals.

B. Principal Component Analysis

Principal Component Analysis (PCA) was also applied to the acoustic features. In the English dataset, the first two principal components accounted for 20.1% and 12.8% of the total variance, respectively. PC1 was mainly associated with voice quality and prosodic energy features, as reflected by the strong influence of loudness-related and F0 features, whereas PC2 was associated with spectral–formant characteristics (*i.e.*, HNR, F3 frequency, shimmer, and F1–F2 amplitude ratios). In the French dataset, the first two principal components accounted for 23.1% and 11.7%. PC1 showed a pattern similar to that in English, being mainly influenced by loudness, spectral balance (*e.g.* alpha ratio, Hammarberg Index), and pitch, whereas PC2 was influenced by formant amplitude and frequency parameters (*e.g.* F1–F3 amplitudes and frequencies). This reveals an overall similar structure between the datasets, with subtle linguistic differences in acoustic information. Furthermore, in both datasets, most of the relevant information was captured in the first 19 components, accounting for about 88% of the total variance. Moreover, 90% was reached with 23 components in English and 21 in French. Based on these results, the feature space was reduced to 19 representative variables per language, preserving most of the variance while minimizing redundancy among correlated features.

C. Acoustic similarity between emotions

In addition, the acoustic similarity between emotions was examined to determine whether some are closely related, which could explain perceptual confusions. To this end, Euclidean distances between the centroid representations of each emotion were computed. The similarities among emotions were then analyzed using hierarchical clustering and Classical Multidimensional Scaling (MDS) to visualize relationships among emotion categories. Mean inter-emotion distances indicated a moderate level of acoustic separability (EN = 2.40 ± 0.82; FR = 2.92 ± 0.99), suggesting a partial overlap between emotions. In both languages, self-confidence, disgust,

and contempt formed consistently close clusters, while anger versus surprise, and sadness versus empathy showed the most significant differences, reflecting distinct acoustic profiles. The proximity of self-confidence, disgust, and contempt in both French and English datasets can induce frequent perceptual confusions. This reflected that self-confidence and contempt, two complex emotions considered in the EVE corpus, may be linked to disgust, as complex emotions are composed of multiple basic affective states. These observations will inform future analyses examining the grouping of acoustically similar emotions and their influence on model performance. In summary, these analyses identified redundant features and acoustically similar emotions, which will inform both feature selection and emotion labeling strategy.

VI. Emotion Recognition methods

Several supervised machine learning models were trained and compared to predict emotion perception from acoustic cues. Baseline regression approaches, such as Ridge Regression, Random Forests (RFs), k-Nearest Neighbors (KNNs), and Linear Support Vector Regression (SVR), were first implemented to establish performance benchmarks. As the system's objective is to predict distributions over emotions rather than a single emotion, regression was preferred over classification methods to reflect the inherent variability of human perception. More complex architectures, such as Multilayer Perceptrons (MLPs), will then be explored given their ability to capture nonlinear relationships between acoustic features and emotions. A more advanced transformer-based approach will also be tested, relying on the HuBERT model [1].

For both languages, the dataset was split into training (70%), validation (15%), and test (15%) subsets. An actor-independent split ensured generalization to unseen speakers, with an even distribution across emotions. Feature normalization was applied using training data to prevent data leakage.

Model performance will be assessed using mean squared error and cosine similarity, computed on predicted emotion distributions to evaluate how closely they approach human perception (*i.e.*, soft labels). The percentage of correctly predicted emotions will also be calculated to measure how often the emotion with the highest predicted probability matches the one most frequently reported by participants. To ensure comparability across label formulations, all MLPs will be evaluated on an identical test set, using the raw-votes label type. All models were initially trained on the complete set of 62 acoustic features, separately for French and English. The reduced feature sets identified through PCA and correlation analyses will later be used to improve model efficiency. An emotion-grouping approach will also be considered to group emotions that are likely to be mistaken for one another. The MLP architecture will be optimized by varying the number of hidden layers and the number of neurons per layer. All models will be implemented in Python using the PyTorch library.

In short, model performance will be analyzed across languages, labeling approaches, and various feature sets to obtain an effective SER system. These analyses are ongoing.

VII. Discussion and conclusion

The proposed soft-label approach offers a realistic way to replicate human emotion perception. Rather than associating each recording with a single emotion, the model aims to predict probability distributions that reflect the inherent uncertainty and variability. This is particularly relevant for VR-based public speaking training systems, where emotion prediction can drive the real-time behavior of embodied virtual audiences. By enabling responsive and realistic audience reactions, this approach supports immersion, social presence, and training effectiveness [6].

Several directions will be pursued to extend this study. Feature reduction techniques will be explored to determine whether smaller feature sets can achieve comparable or improved performance. In parallel, emotion grouping strategies will be investigated from both acoustic and perceptual perspectives (*i.e.*, based on frequently confused emotions in the EVE corpus) [10]. As the SER system targets both French- and English-speaking users, multilingual modeling will also be considered. Model architectures will be further refined to include advanced approaches, such as HuBERT. Finally, the model will be assessed across multiple corpora, including a comparison between acted and perceived emotion annotations within the EVE dataset.

The MLP trained on soft labels with confidence weighting is expected to achieve the best performance, as this strategy reduces annotation noise by weighting emotion judgments according to listener confidence and, thus, their reliability. Deep learning architectures should also better capture the complex relationship between acoustic features and emotions. However, it will be necessary to check whether such improvements reflect better predictions or result from smoother distribution shapes. Performances are expected to be consistent across French and English, supporting future work on multilingual SER models.

While feature pruning has been completed, extensive work remains to develop models that are sufficiently accurate and fast for real-time use in VR applications. Several VR public speaking environments have already been created, simulating a range of situations (e.g., classroom, courtroom, job interview). Therefore, the next step is to integrate the proposed SER system into these environments to drive virtual audiences' behavior. By mapping the predicted emotion probability distributions to virtual agents, the audience will be able to respond dynamically and realistically to the speaker's emotions. Such emotionally responsive audiences are expected to enhance realism, immersion, and social presence, while also allowing the audience's behavior to influence the speaker's emotional state [6] [26], as in real public speaking situations. This interaction is essential for creating training environments that effectively prepare users for real-world speaking scenarios. Although this constitutes a primary application, the approach could be extended to other interactive systems, such as virtual assistants or social VR applications requiring emotionally adaptive behavior.

REFERENCES

[1] M. Abdelrahman, H. Wei-Ning, and L. Kushal, "HuBERT: Self-supervised representation learning for speech recognition, generation, and compression," Meta. [Online]. Available: https://ai.meta.com/blog/hubert-self-supervised-representation-learning-for-speech-recognition-generation-and-compression/. Accessed: Oct. 20, 2025.

[2] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Dec. 2019, doi: 10.1016/j.specom.2019.12.001.

[3] M. Bachmann, A. Subramaniam, J. Born, and D. Weibel, "Virtual reality public speaking training: Effectiveness and user technology acceptance," *Frontiers in Virtual Reality*, vol. 4, Sep. 2023, doi: 10.3389/frvir.2023.1242544.

[4] T. Bänziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech Communication*, vol. 46, no. 3–4, pp. 252–267, May 2005, doi: 10.1016/j.specom.2005.02.016.

[5] A. Chakhtouna, S. Sekkate, and A. Adib, "Unveiling embedded features in Wav2Vec2 and HuBERT models for speech emotion recognition," *Procedia Computer Science*, vol. 232, pp. 2560–2569, Jan. 2024, doi: 10.1016/j.procs.2024.02.074.

[6] M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro, and S. Scherer, "Exploring feedback strategies to improve public speaking: An interactive virtual audience framework," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*, 2015, pp. 1143–1154, doi: 10.1145/2750858.2806060.

[7] W. T. Coombs and E. R. Tachkova, "How emotions can enhance crisis communication: Theorizing around moral outrage," *Journal of Public Relations Research*, vol. 36, no. 1, pp. 6–22, Aug. 2023, doi: 10.1080/1062726X.2023.2244615.

[8] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3–4, pp. 169–200, May 1992, doi: 10.1080/02699939208411068.

[9] M. El-Yamri, A. Romero-Hernandez, M. Gonzalez-Riojo, and B. Manero, "Emotions-responsive audiences for VR public speaking simulators based on the speakers' voice," in *Proceedings of the IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, Maceio, Brazil, 2019, pp. 349–353, doi: 10.1109/ICALT.2019.00108.

[10] E. Etienne, A. Remacle, A.-L. Leclercq, and M. Schyns, "EVE: Emotional validated expressions, an acted audiovisual corpus," in *Proceedings of the 25th ACM Int. Conf. Intelligent Virtual Agents (IVA '25)*, Berlin, Germany, 2025, doi: 10.1145/3717511.3749303.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Jul. 2015, doi: 10.1109/TAFFC.2015.2457417.

[12] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, Nov. 2007, doi: 10.1111/j.1467-9280.2007.02024.x.

[13] A. Goldenberg, E. Weisz, T. D. Sweeny, M. Cikara, and J. J. Gross, "The crowd-emotion-amplification effect," *Psychological Science*, vol. 32, no. 3, pp. 437–450, Feb. 2021, doi: 10.1177/0956797620970561.

[14] J. Y. Han, B. R. Shaw, R. P. Hawkins, S. Pingree, F. McTavish, and D. H. Gustafson, "Expressing positive emotions within online support groups by women with breast cancer," *Journal of Health Psychology*, vol. 13, no. 8, pp. 1002–1007, Nov. 2008, doi: 10.1177/1359105308097963.

[15] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech emotion recognition using deep neural networks considering verbal and nonverbal speech sounds," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2019, doi: 10.1109/ICASSP.2019.8682283.

[16] N. Kamaruddin, A. Wahab, and C. Quek, "Cultural dependency analysis for understanding speech emotion," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5115–5133, Nov. 2011, doi: 10.1016/j.eswa.2011.11.028.

[17] S. C. Leong, Y. M. Tang, C. H. Lai, and C. K. M. Lee, "Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing," *Computer Science Review*, vol. 48, p. 100545, Feb. 2023, doi: 10.1016/j.cosrev.2023.100545.

[18] P. D. MacIntyre and J. R. MacDonald, "Public speaking anxiety: Perceived competence and audience congeniality," *Communication Education*, vol. 47, no. 4, pp. 359–365, Oct. 1998, doi: 10.1080/03634529809379142.

[19] P. D. MacIntyre, K. A. Thivierge, and J. R. MacDonald, "The effects of audience interest, responsiveness, and evaluation on public speaking anxiety and related variables," *Communication Research Reports*, vol. 14, no. 2, pp. 157–168, Mar. 1997, doi: 10.1080/08824099709388657.

[20] J. P. Mulki, F. Jaramillo, E. A. Goad, and M. R. Pesquera, "Regulation of emotions, interpersonal conflict, and job performance for salespeople," *Journal of Business Research*, vol. 68, no. 3, pp. 623–630, Sep. 2014, doi: 10.1016/j.jbusres.2014.08.009.

[21] F. Palmas, J. Cichor, D. A. Plecher, and G. Klinker, "Acceptance and effectiveness of a virtual reality public speaking training," in *Proceedings of the 2019 IEEE International Symposium Mixed and Augmented Reality (ISMAR)*, Oct. 2019, pp. 363–371.

[22] R. Rajoo and C. C. Aun, "Influences of languages in speech emotion recognition: A comparative study using Malay, English, and Mandarin languages," in *Proceedings og the IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, Penang, Malaysia, 2016, pp. 35–39, doi: 10.1109/ISCAIE.2016.7575033.

[23] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: A survey," *Social Network Analysis and Mining*, vol. 8, no. 1, Apr. 2018, doi: 10.1007/s13278-018-0505-2.

[24] S. Saufnay, E. Etienne, and M. Schyns, "Improvement of public speaking skills using virtual reality: Development of a training system," in *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Glasgow, United Kingdom, 2024, pp. 122–124, doi: 10.1109/ACIIW63320.2024.00025.

[25] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, Jun. 2018, doi: 10.3390/s18072074.

[26] M. Slater, D. P. Pertaub, and A. Steed, "Public speaking in virtual reality: Facing an audience of avatars," *IEEE Computer Graphics and Applications*, vol. 19, no. 2, pp. 6–9, Apr. 1999.

[27] M. Takac, J. Collett, K. J. Blom, R. Conduit, I. Rehm, and A. De Foe, "Public speaking anxiety decreases within repeated virtual reality training sessions," *PLOS ONE*, vol. 14, no. 5, p. e0216288, May 2019.

[28] P. Totterdell and D. Holman, "Emotion regulation in customer service roles: Testing a model of emotional labor," *Journal of Occupational Health Psychology*, vol. 8, no. 1, pp. 55–73, Jan. 2003, doi: 10.1037/1076-8998.8.1.55.

[29] J. L. Tracy and D. Randles, "Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt," *Emotion Review*, vol. 3, no. 4, pp. 397–405, Sep. 2011, doi: 10.1177/1754073911410747.

[30] G. K. Verma and U. S. Tiwary, "Affect representation and recognition in 3D continuous valence–arousal–dominance space," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2159–2183, Jan. 2016, doi: 10.1007/s11042-015-3119-y.

[31] S. Weilenmann, U. Schnyder, B. Parkinson, C. Corda, R. Von Känel, and M. C. Pfaltz, "Emotion transfer, emotion regulation, and empathy-related processes in physician–patient interactions and their association with physician well-being: A theoretical model," *Frontiers in Psychiatry*, vol. 9, Aug. 2018, doi: 10.3389/fpsyt.2018.00389.

[32] Y. Xiao, L. Hudders, A.-S. Claeys, and V. Cauberghe, "The impact of expressing mixed valence emotions in organizational crisis communication on consumers' negative word-of-mouth intention," *Public Relations Review*, vol. 44, no. 5, pp. 794–806, Oct. 2018, doi: 10.1016/j.pubrev.2018.10.007.

[33] D. Zapf, "Emotion work and psychological well-being," *Human Resource Management Review*, vol. 12, no. 2, pp. 237–268, Jun. 2002, doi: 10.1016/S1053-4822(02)00048-7.