11. EmoVox: Creation of a speech database for emotion analysis

Etienne, E.[a], Leclercq, A.-L.[bcd], Remacle, A.[b], and Schyns M.[a]

[a] *QuantOM , HEC Liège, University of Liège*
[b] *Département de Logopédie, Université de Liège*
[c] *Unité de Recherche Enfances*
[d] *Clinique Psychologique et Logopédique de l'Université de Liège*

Type of manuscript: Extended abstract

Keywords: speech database, emotions, training, deep learning

Extended abstract

This work is based on a doctoral thesis in progress.

Public speaking is prevalent in various fields, particularly in business and marketing, encompassing activities such as pitching products or services, delivering progress reports, presenting proposals, and interacting with customers. Training in public speaking is of paramount importance, yet providing effective training can be challenging. The Metaverse and Virtual reality (VR), in general, have emerged as promising technologies for training, offering an immersive and interactive learning experience that enhances skills, knowledge, and confidence when interacting with others. In this context, emotions play a significant role, necessitating an automated decision-support solution to improve participants' performance according to their emotional state (e.g., joy, sadness, anger). Accurate detection of emotions allows virtual avatars in the training environment to react appropriately to the speech and emotional state of the speaker, increasing the sense of presence and intervention effectiveness. Furthermore, it will enhance communication in the Metaverse (Daneshfar and Jamshidi, 2023). However, the importance of emotion detection goes beyond Virtual Reality and can have many benefits when dealing with robots, chatbots, virtual assistants and smart speakers in general. Artificial Intelligence (AI) techniques, such as machine learning and deep learning, can be employed to detect emotions effectively. Data is crucial to the development and effectiveness of AI algorithms (Abbaschian *et al.*, 2021; Lieskovská *et al.*, 2021). Indeed, the algorithms are designed to learn from data, recognise patterns, and make predictions based on that data. Without sufficient and high-quality data, AI models cannot accurately learn and make accurate predictions. Good oral production databases with thousands of sound data labelled according to emotions are thus required.

This research aims to develop a high-quality emotional speech dataset in French and English and validate the dataset through subjective evaluation experiments assessing human accuracy in recognising the intended emotion.

The motivation for creating the database is the need for a publicly available high-quality emotion recognition database. Existing databases, such as IEMOCAP (Busso *et al.*, 2008) or CREMAD (Cao *et al.*, 2014), do not contain sufficient data for deep learning approaches. Hume's databases (https://hume.ai/solutions) could be utilised, but their limitations include the nature of the proposed texts and their costs. Moreover, these databases are primarily in English. Although several French databases exist, such as The French Emotional Speech Database – Oréau (Kerkeni *et al.*, 2020) or the Canadian French Emotional speech dataset (Gournay *et al.*, 2018), a Canadian database in French), they are too small for machine learning exploitation.

In this paper, emotions are treated as discrete states (categorical model), utilising the six basic Ekman emotions (Ekman, 2013) and other common emotions during public speaking (These corpora contain 72 lists of 10 phonetically balanced sentences. We will use one of the lists. We will ask each speaker to record ten sentences with the different emotions (6 basic Ekman emotions (anger, disgust, joy/enthusiasm, fear, surprise sadness) as well as four others useful in the context of public speaking (anxiety/embarrassment, confusion, contempt, shyness). The research methodology comprises two studies. The first involves creating audio recordings using either professional actors (approximately sixty French native speakers and 60 English native speakers) or Text-to-Speech software capable of generating emotional samples (e.g., https://www.texttovoice.online/). The second study is a subjective evaluation experiment assessing human accuracy in recognising the intended emotion.

For the first study, actors will be recruited from French-speaking regions in Belgium or England. Sentences will be derived from phonetically balanced corpora, such as the *Harvard Sentences* in English (Rothauser, 1969) and the *Fhavard corpus* in French (Aubanel, 2020). These corpora contain seventy-two lists of ten phonetically balanced sentences. We will use one of the lists. We will ask each speaker to record ten sentences with the different. Each speaker will record ten sentences with different emotions, and each sentence will be interpreted by twelve different speakers (Gournay *et al.*, 2018). Audio recordings based on Text-to-Speech software will be created using a website API. The second study involves validating the emotion in each audio recording through evaluations by 25 judges (determined based on a power analysis). These judges, naive listeners participating in citizen science, will listen to the recordings and assign emotional valence. Each participant will judge 90 statements, evaluating the emotion in the audio clip by selecting from the ten proposed emotions. Participants will listen to audio clips using headphones and answer numerical scales to judge the chosen emotion's intensity and evaluate the performance's artificial level. They will also complete a questionnaire with socio-demographic data.

Sentences created with the Text-to-Speech software are already available. Actor recordings will begin in April, with the validation of the first recording to be completed by the end of April. Mid-May expects preliminary results.

This paper is original in its aim to create and validate an emotional speech dataset in two phases for use in VR public speaking training environments. It addresses the limitations of existing datasets by providing a high-quality dataset in both French and English, and it includes a validation process involving subjective evaluation experiments.

References

Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4), 1249.

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.

Buechel, S., & Hahn, U. (2022). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

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.

Daneshfar, F., & Jamshidi, M. B. (2023). An octonion-based nonlinear echo state network for speech emotion recognition in Metaverse. Neural Networks, 163, 108-121.Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings* (Vol. 11). Elsevier.

Kerkeni, L. , Cleder, C., Serrestou, Y. & Raoof, K. (2020). French Emotional Speech Database - Oréau (Version 1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4405783

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, *10*(10), 1163.

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

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).

Rothauser, E. H. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225-246.