The development of AI has created sophisticated applications, but they often lack the ability to understand non-verbal cues essential for human communication. This thesis addresses that gap by exploring audio-visual deep learning methods for Affective Computing.

We believe that analysing deep learning model behavior can improve performance and build trust by making their decision-making processes more transparent. Our work focuses on processing voice signals and facial expressions, specifically smiles and laughter, as key indicators of emotional states.

The main contributions of this research are fourfold. First, we enhanced three existing datasets by adding annotations for speaker/listener roles and the intensity of smiles and laughter. Then, using LSN-TCN, a deep learning-based neural network, we analyzed how fusing audio and visual feature representations impacts the detection of smiles and laughter. We also implemented Social-MAE, an advanced multimodal system that effectively encodes facial and vocal information for tasks like emotion recognition. Finally, we explored a novel method to separate affective information from existing deep learning systems without compromising their performance by using an auxiliary network.

This thesis provides open-source methods to leverage non-verbal cues, paving the way for more sophisticated and empathetic AI systems with potential applications in social and clinical settings.