Emotion detection from speech in dialog systems: advances and challenges

Stefano Rovetta[1], Zied Mnasri[1,2], Francesco Masulli[1], Alberto Cabri[1]

{Stefano.rovetta, francesco.masulli}@unige.it

{Zied.mnasri, alberto.cabri}@dibris.unige.it

[1] DIBRIS, University of Genoa, Genova, Italy

[2] ENIT, University of Tunis El Manar, Tunis, Tunisia

Abstract

I. Interest of the problem

Expressive communication is the next challenge of human-machine interaction systems. Since a few years, machines have been able to recognize and synthesize human speech with a high level of quality; however, it is still problematic to achieve effective *artificial empathy*, where the machine would be able to detect human emotion online, in order to provide a consequent reaction.

II. State of the art

Offline emotion recognition from speech has reached a good level, especially using supervised learning models, such as DNN, RNN and CNN, using either standard features (GeMAPS), spectrogram images, or raw audio data. However, the main criticism is about the inherent overfitting that makes such models less efficient when applied on real-world data. This weakness was revealed by cross-corpus tests, where different corpora, often from different languages and different labeling systems, are used to validate and test the trained models. Another approach consists in aggregating such different corpora to train the model on heterogeneous data. To cope with the different labeling systems, some meta-labels, such as low/high valence, arousal or dominance have been used instead of explicit emotion categories. Nevertheless, both approaches confirmed the overfitting tendency of supervised models, since their respective performances have been usually lower when tested on datasets extracted from different corpora.

III. Proposed method

Emotion understanding via *self-organized* machine learning, for instance by clustering, could provide an effective change of perspective. The recent advances in soft clustering make it possible to discover and analyze latent data structures. Experiments show that unsupervised methods achieve the same performance as supervised classifiers on cross-corpus data, which provides a major reason to opt for clustering, since it is much less expensive in terms of cost, time and errors.

IV. Challenges and applications

Clustering models are aimed to detect emotion and/or emotion change online in a dialog scenario, so that the machine can adapt its reaction to the detected emotion. Applications may include affective communication with humanoid robots, customer mood detection by virtual purchase assistants, monitoring or diagnosing degenerative diseases like Parkinson, Alzheimer and multiple sclerosis which are related to rapid change in mood.

Keywords- Emotion detection from speech, dialog systems, human-machine interaction, clustering.