ABSTRACT

Humans perceive emotion in multimodal ways. Speech is one of the sensory modalities in which emotions can be perceived. Within speech, humans communicate emotion through acoustic and linguistic information. In automatic emotion recognition by computers, known as affective computing, there is a shift from unimodal acoustic analysis to multimodal information fusion. As in human speech emotion perception, computers should be able to perform speech emotion recognition (SER) from bimodal acoustic-linguistic information fusion.

This research aims to investigate the necessity to fuse acoustic with linguistic information for recognizing dimensional emotions. To achieve this goal, three sub-goals were addressed: SER by using acoustic features only, fusing acoustic and linguistic information at the feature level, and fusing acoustic and linguistic information at the decision level.

The first strategy aims at maximizing the potency of recognizing dimensional SER by merely using acoustic information through investigating the region of analysis and the effect of silent pause regions. This study generalizes the effectiveness of means and standard deviations to represent acoustic features and the prediction of the importance of silent pause regions for dimensional SER. In addition, the aggregation of acoustic feature models valence and arousal prediction better than the majority voting method. Although several approaches have been carried out, acoustic-based dimensional SER still has some limitations. The major drawback is the low performance of valence's prediction score.

The second and third strategies aim at improving the valence prediction, investigating the necessity of bimodal information fusion, and evaluating the fusion frameworks for fusing acoustic and linguistic information. Two fusion methods for acoustic-linguistic information fusion are studied namely early fusion approach and late-fusion approach. At the feature level (FL) or early fusion approach, two fusion methods are evaluated --- feature concatenation and network concatenation. The FL methods showed significant performance improvements over unimodal dimensional SER. At the decision level (DL) or late-fusion approach, acoustic and linguistic information are trained independently, and the results are fused by support vector machine (SVM) to make the final predictions. Although this proposal is more complex than the previous FL fusion, the results showed improvements over the previous DL approach. These studies revealed the necessity to fuse acoustic with linguistic features for dimensional SER.

This study links the current problems in dimensional SER with its potential solutions. The fusion of acoustic and linguistic information fills the gap in dimensional SER. The FL approach improved the performance of unimodal SER significantly. The DL approach improves the FL approach's performance by fusing decisions obtained from the bimodal FL approaches. The results of this research are expected to contribute to gaining better insights for the future strategy in implementing SER, whether to use acoustic-only features (less complex, less accurate), early fusion method (more complex, more accurate), or late-fusion method (most complex, most accurate).

Keywords: dimensional emotion, affective computing, speech emotion recognition, information fusion, acoustic information