Emotion Intensity and Gender Detection via Speech and Facial Expressions

Elahe Bagheri, Oliver Roesler, Hoang-Long Cao and Bram Vanderborght

Introduction

Detecting human emotions is crucial in developing cognitive and adaptive behaviors for artificial intelligent systems, robots, and (virtual) agents. Emotion detection is the ability to recognize another's affective state, which typically involves the integration and analysis of human expressions through different modalities, like facial expression, speech, body movements, and gestures. Researches showed 55% of human emotions convey through facial expression and 38% through speech, therefore, facial and speech emotion recognition received significant attention during the last decades. Although finding the type of expressed emotion is essential to adapt to a user's affective state, it is not enough, and a difference in intensity has been proven to be important to distinguish different emotional states. For instance, a polite smile versus an embarrassed smile and posed versus spontaneous smile are separable by differences in their expression intensities.

Emotion Intensity detection

The highest accuracies for emotion intensity detection obtained by Model3 which consists of CNN, BiLSTM and Attention layers as is shown in following table. The facial and speech features of six basic emotions expressed in RAVDESS dataset are used for this task. While using facial features lead to higher accurcy for males, speech features led to higher accuracy for females.

Data	Facial Features	Speech Feature
Female & Male	56.24	73.53
Female	54.72	75.67
Male	58.31	65.5

The proposed model could outperform the state-of-the-art in emotion intensity detection based on speech features as is shown in the following table.

Research	Architecture	Acc
Jalal [1]	CNN+BiLSTM+CapNet	70.4%
Model3	CNN+BiLSTM+Att	**73.53%**

Gender Detection

Since the obtained accuracies in emotion intensity detection for males and females are noticeably different, in this experiment we investigate speech and facial features for the task of gender detection.

Model3 obtained an accuracy of 89.8% for gender detection using the MFCC and PMC feature sets of a speech signal, which is the highest obtained accuracy in comparison with the other models.

A gender detection model should be robust to various emotions, however, different state-of-the-art models considered a different number of emotion classes. For instance, Shaqra et al. [2] considered six emotions and obtained an accuracy of 98.67%. Although the proposed model could not beat the state-of-the-art, it is more robust since considers more emotional states, i.e., eight classes.

Research	Acc
Shaqra et al.[2]	98.67%
Proposed model	89.8%

Materials & Methods

The following table shows the number, type, and order of layers of the proposed architectures. The parameter settings are as follows: 1D convolution layers with 64 filters and a kernel size of 3 are used. ReLU activation function is applied for adding non-linearity. Dropout layers are used as regularizers with a ratio of 0.1. 1D max-pooling layers with a kernel size of 4 are used to introduce sparsity in the network parameters. Dense layers with the activation functions of sigmoid are used. The number of epochs and batch size are set to 250 and 128. LSTM and BiLSTM layers have five units.

Model1	Model2	Model3
CNN	CNN	CNN
CNN	CNN	MaxPooling
Dropout	MaxPooling	CNN
MaxPooling	Flatten	MaxPooling
Flatten	Bi/LSTM	Flatten
Dense	Dense	Bi/LSTM
Dense		Attention
		Dropout
		Dense

Discussion

The obtained results show using facial features leads to more accurate emotion intensity detection for males in comparison to females, while using speech features lead to higher accuracy in females' emotion intensity detection.

Since some of the speech features related to emotion recognition are related to the subject's gender, we hypothesized that females convey more details about the intensity of their emotions through their speech. To verify this hypothesis, model 3 is applied to both the facial and speech features of each individual subject, where the minimum and maximum accuracies for males are 63.92% and 78.79%, respectively, while the corresponding values for females are 58.26% and 71.15%. On the other hand, the minimum and maximum accuracies for emotion intensity detection via speech features for females are 71.49% and 95.83% while the corresponding values for males are 59.29% and 85.71%, respectively.

Conclusion

In this study, we designed different deep neural network-based architectures for emotion intensity and gender detection using facial and speech features obtained from open-source toolkits.

The RAVDESS dataset was used to evaluate the proposed architectures because it categorizes emotions based on their intensity.

The results showed speech features lead to higher accuracy in emotion intensity detection than facial features, and in the RAVDESS dataset females convey more details about the intensity of their emotions through their speech than males, however, in the absence of speech signal, emotion intensity detection is more accurate for males than females, i.e., males convey more details about the intensity of their emotions through their face.

Further, we used the proposed model for gender detection using facial and speech features. The obtained results show that gender detection is also more accurate for speech than facial features for the RAVDESS dataset and that the proposed model is comparable with the state-of-the-art while it is more robust, i.e., handles more emotional states.

References

[1] Md Asif Jalal, Erfan Loweimi, Roger K Moore, and Thomas Hain. Learning temporal clusters using capsule routing for speech emotion recognition. In *Proc. Interspeech*, pages 1701–1705, 2019.

[2] Ftoon Abu Shaqra, Rehab Duwairi, and Mahmoud Al-Ayyoub. Recognizing emotion from speech based on age and gender using hierarchical models. *Procedia Computer Science*, 151:37–44, 2019.

Future Research

In future work, we are going to use the proposed model in an empathy framework, which consists of a facial emotion detection model for emotion type detection. The framework will be used in a Human-Robotic Interaction scenario, where a robot adjusts its behaviors and applies empathy based on users' emotion type and intensity. The robot uses the proposed model to detect users' emotional intensity and regulate its reactions. As empathy level differs based on the intensity of the user's emotion, if the user's emotional intensity is low the robot's empathic behavior is limited to expressing similar emotions with those that the user is experiencing (parallel empathy), while if it has high intensity the robot shows higher levels of empathy (reactive empathy).

Contact Information

Email elahe.bagheri@vub.be

Address Robotics and Multibody Mechanics Research Group, Vrije Universiteit Brussel, Brussels, Belgium.