Q1. The features that are manually engineered by a person are called hand engineered features Some examples are prosodic features (pitch, energy, zero-crossings), spectral features, non-linear features like Teager energy operator (TEO), etc. On the other hand, learned features are ones that are obtained/learned by a machine learning algorithm basically high level abstractions obtained from raw data.

Q2.

The diverse results suggest that we are not yet at a stage where we could afford to cast away hand-crafted features. Especially, in tasks with sparse data it may still be safer to rely on conventional feature sets. While learning from raw audio outperformed hand-crafted features on the emotional corpus, it performed worse on the smaller crying database. However, the difference disappeared when the layers were pretrained on spectrograms we had extracted from another dataset. This shows that feature engineering can still help improve the robustness of end-to-end systems. It is difficult to master hand-crafted feature engineering due to the varieties of data and each requiring different features engineering techniques. But with the help of domain expertise, engineering the best suited features for a specific task would make the model perform. When it comes to learned representations, analyzing them is a difficult task as neural networks are like a black box as compared to hand crafted features which can easily be understood and analyzed.

Q3.

On the datasets involving atypical/emotional speech, end-to-end speech processing system can perform efficiently. This efficiency is due the fact that the end-to-end speech processing system learns the data representation directly from the raw waveforms and large amount of dataset makes models learn and generalize well.