Speech Emotion Recognition

Gaurav Sahu, MMath CS, University of Waterloo

February 8, 2019

Problem Statement

The aim of this project is to explore the domain of Speech Emotion Recognition (SER), i.e., automatically identifying human emotions from speech. It would be specifically helpful to use speech as it provides us with additional information that could aid ambiguity-resolution in communication. For example, "I'm fine", could be said in multiple ways and we would not be able to identify the correct context if we only consider textual features. However, the task of SER is not trivial as the idea of "emotion" is innately quite subjective and it could be a bit challenging for a machine to understand. Most of the works in literature focus on the "big six" emotion-classes, namely, anger, disgust, fear, happiness, sadness, and neutral. In this work, we would be using the same and would especially focus on identifying emotion indicators using audio-based features such as pitch, loudness, and pause. More details about the approach and the dataset to be employed are described in the following sections.

Dataset

We would be using the following datasets for our experiments:

1. *IEMOCAP:* The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) dataset [1] was released in 2008 by researchers in University of Southern California (USC). It contains five recorded sessions of conversations from ten speakers and amounts to 12 hours of audio-visual information. It is annotated with categorical emotion labels (Anger, Happiness, Sadness, Neutral, Surprise, Fear, Frustration and Excited) and dimensional labels (values of the activation and valence from 1 to 5).
2. *Emotional Prosody Speech and Transcripts:* The dataset [2] was released in 2002 and contains recordings of professional actors reading a series of semantically neutral utterances (dates and numbers) spanning fourteen distinct emotional categories. We shall reduce the 14 categories to "big six" for our experiments.

Approach

The task of SER is not new and has been studied for quite some time in literature. Recent introduction of deep neural networks to the domain has also significantly improved the state-of-the-art performance. In this work, we will construct features from audio alone and observe what effects does it have when combined with textual features. We now describe some relevant features from the literature [3] for our *audio-only* framework:

1. **Pitch:** Pitch is important because wave forms produced by our vocal cords change depending on our emotion. Many algorithms for estimating the pitch signal exist. We will use the most common method based on *autocorrelation of center-clipped* frames [4].
2. **Number of Harmonics:** In the emotional state of anger or for stressed speech, there are additional excitation signals other that pitch ([5], [6]). This additional excitation is apparent in the spectrum as harmonics and cross-harmonics. We will use the Teager Energy Opertor [5] to calculate this feature.
3. **Pause:** This feature represents the "silent" portion in the audio data. This quantity is directly related to our emotions; For example, we tend to speak very fast (resulting in low value of the feature) when excited (say, angry or happy).
4. **Speech Energy:** This feature is related to the loudness in speech and hence, can be used for emotion recognition. We will use the standard energy formula for a signal [7] to calculate this feature.

Once we extract features from the audio sample, we can train a multi-class SVM classifier to guess the correct emotion or we could also employ a clustering technique such as K-means.

We will compare the performance of *audio-only* model with a *text-only* seq2seq model with Bi-LSTM [8] cells (current state-of-the-art for text sentiment analysis [9]) to get a better understanding of the relevance of audio features. Both the datasets mentioned earlier contain audio transcripts; However, we will only use transcripts from the IEMOCAP dataset while building the *text-only* baseline model for the latter contains semantically neutral text.

Finally, we will compare our model's performance with other state-of-the-art methods for emotion recognition which combine features from multiple modalities using deep learning.

References

[1] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[2] M. Liberman, "Emotional prosody speech and transcripts," *http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp? catalogId= LDC2002S28*, 2002.

[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[4] M. Sondhi, "New methods of pitch extraction," *IEEE Transactions on audio and electroacoustics*, vol. 16, no. 2, pp. 262–266, 1968.

[5] H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech production and speech modelling*, pp. 241–261, Springer, 1990.

[6] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 9, no. 3, pp. 201–216, 2001.

[7] "Energy (signal processing)," *Wikipedia*.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] A. Radford, R. Jozefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," *arXiv preprint arXiv:1704.01444*, 2017.