Speech Emotion Recognition

Gaurav Sahu, MMath CS, University of Waterloo

April 5, 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 and Preprocessing

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. *SAVEE:* The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset [2] was released in 2014 by researchers at the University of Surrey and consists of recordings from 4 male actors in 7 different emotions (anger, disgust, fear, happiness, sadness, neutral and surprise), 480 British English utterances in total. It also comes with text annotations with 15 sentences per emotion: 3 common, 2 emotion-specific and 10 generic.

The SAVEE dataset was straightforward to deal with as it contained short wav files with their respective annotations. However, the IEMOCAP dataset provides wav files for individual sessions. So each of the session-specific wav files were segmented into shorter files in order to construct a corpus aptly aligned with its annotations. Finally, all the wav files were trimmed to remove the *silent noise* from the audio.

Proposed Approach and Algorithms

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 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 and respective algorithms to compute them:

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].

   Formally, the input signal $y[n]$ is center-clipped to give a resultant signal, $y_{clipped}[n]$, given by:

   $$y_{clipped}[n] = \begin{cases} y[n] - C_l, & \text{if } y[n] \geq C_l \\ 0, & \text{if } |y[n]| < C_l \\ y[n] + C_l, & \text{if } y[n] \leq C_l \end{cases}$$

   Here, C_l is nearly half the mean of the input signal and $[\cdot]$ denotes that the input signal is discrete in nature. Now, autocorrelation is calculated for the obtained signal $y_{clipped}$, which is further normalized and finally the peak values are taken and associated with the pitch of the given input $y[n]$. It was found that center-clipping the input signal resulted in more distinct autocorrelation peaks.

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 (Also shown by 1(a)). Harmonic calculation is based on the algorithm described in [7]

3. **Speech Energy:** This feature is related to the loudness in speech and hence, can be used for emotion recognition (See 1 (b) for reference). We use the standard Root Mean Square Energy (RMSE) to calculate the signal given by:

   $$E = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y[i]^2} \tag{1}$$

4. **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). In order to calculate this value, we set a manual threshold $t = 0.4 * E$ and the feature value is give by the probability $P(y[n]) < t$.

Apart from the features described above, some other statistical metrics such as mean and standard deviation of the input signal have been used as features too.

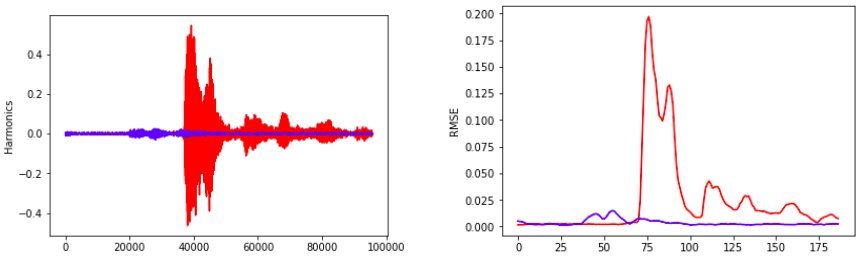
All the pre-processing

Evaluation Measures and Next Steps

Now that we have extracted features from audio samples, the next steps would be to:

1. Train a multi-class classifier (such as SVM, Random Forest, and Gradient Boosting)
2. Implement the *text-only* baseline sequence-to-sequence architecture [8] with attention mechanism [9] and evaluate it against the available transcriptions
3. Combine features from both text and audio and observe the effect in performance
4. Implement the current state-of-the-art for multi-modal sentiment classification [10] and compare its performance with that of the proposed model's

We will use classification accuracy as the primary criterion for evaluation; However, accuracy could be sometimes misleading and hence we would also calculate a more normalized metric, the F-score, which is defined the harmonic mean of precision and recall.



(a) Harmonics of angry (red) and sad (blue) audio signals

(b) RMSE plots of angry (red) and sad (blue) audio signals

Figure 1: Comparison of two "opposing" audio signals

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] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[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] D. Fitzgerald, "Harmonic/percussive separation using median filtering," 2010.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[10] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

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

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