Analysis of Speech Emotions in Realistic Environments

Biswajit Dev Sarma[1], *Rohan Kumar Das*[3], *Abhishek Dey*[1] *and Risto Haukioja*[2]

[1]Bay Area Advanced Analytics India (P) Ltd., a Kaliber.AI company, Guwahati, India
[2]Kaliber Labs Inc, San Francisco, USA
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

{biswajit, abhishek, risto}@kaliberlabs.com, rohankd@nus.edu.sg

Abstract

The classification of emotional speech is a challenging task and it depends critically on the correctness of labeled data. Most of the databases used for research purposes are either acted or simulated. Annotation of such acted database is easier as the actor exaggerates the emotions. On the other hand, emotion labeling on real-world data is very difficult due to confusion among the emotion classes. Another problem in such scenario is the class imbalance, because most of the data is found to be neutral in realistic environment. In this study, we perform emotion labeling on realistic data in a customized manner using emotion priority and confidence level. The annotated speech corpus is then used for analysis and study. Percentage distribution of different emotion classes in the real-world data and the confusions between the emotions during labeling are presented.

Index Terms: emotion classification, realistic data, emotion annotation

1. Introduction

The recent growth in the field of speech processing has shown scope for different tasks, emotion recognition being one of them [1–5]. Like all other speech processing applications, quality of speech database plays an important role in speech emotion recognition [6]. It was shown that the recognition performance depends more on the type of data than the choice of input features [7]. Performance of the system in practical condition depends on the degree of naturalness of the database used for training. In most of the openly available speech emotion databases, emotions are simulated by professional or nonprofessional actors [8, 9]. Further, the emotions are not produced in a conversational context leading to lack of naturalness in the emotion content. However, while coming to the real-world applications for emotion recognition, emotion models trained on non-acted speech database is preferred. Because the models trained on acted speech database may have huge variations from that in the real-world scenario.

There comes a lot of challenges while dealing with realistic data. Presence of ambient noise in speech data can severely reduce the performance for emotion classification. The authors of [10] have studied emotions in noisy scenario and found that there is a significant influence of noise in detecting emotions. Further, they proposed a framework with adaptive noise cancellation and few novel features using pitch and energy contours for having improved performance under such conditions. In [11], the authors focused on realistic data, where feature enhancement using neural networks have been introduced for emotion classification. The importance of context information for emotion classification in practical database is demonstrated by the authors of [12]. However, most of the studies available in the literature for emotion classification under adverse conditions are mostly simulated for the studies. Therefore, there is a need to explore realistic data for emotion analysis. In this regard, we focus on the analysis of emotion classes on a database collected from conversational speech in practical settings.

The studies for emotion recognition require appropriate emotion annotation for correct classification. In case of simulated or acted speech, the labeling of emotion classes is easier because the emotion label is predefined or scripted. On the other hand, labeling of practical conversational speech is comparatively difficult and requires careful attention prior to choosing an appropriate label, as the performance of system directly depends on availability of correctly annotated speech corpus. Therefore, annotating process on realistic data becomes pivotal for correct analysis. There exists different tools for the purpose of annotation, which has a marker panel to label the emotions. Wavesurfer is one of the most widely used tool for speech processing applications that can be used for annotating emotion as well [13]. There are few other annotation tools like ANNEMO (ANNotating EMOtions) and Higgins annotation tool that have been designed for annotating emotion from audio-visual data [14, 15]. In this work, for the sake of simplicity to the users Wavesurfer is used for annotating emotions on realistic data.

As discussed the process of annotation on real-world data is very important to have the ground truths to be correct before proceeding with any further study. Therefore, we have customized the process of annotating speech emotion in this study. The users will mark top three emotion classes associated with each speech segment according to the priority based on their perceptual evaluation. Here, top three emotion classes are considered to have a more appropriate label for a segment. Further, they use another panel to provide a confidence score in the range 1 to 5 associated with each emotion class which are marked. The motivation of using this confidence scores for annotating emotion has come from the Geneva emotion wheel (GEW) structure [16]. We believe that priority based top three emotions and confidence score scheme would provide a better estimate of an emotion class for realistic data. Further, as continuous conversational speech is used for the study, a speech segmentation algorithm is introduced to first obtain segments from the continuous speech to annotate emotions. Finally, we present a study based on the annotated corpus that can provide useful remarks for the confusing emotion classes in real-world scenario.

The rest of the work is organized in the following manner. Section 2 details about the realistic database and the process of customized annotation process for speech emotions. The studies and the discussions made on the annotated speech corpus used in this work are reported in Section 3. Section 4 finally provides the conclusion of the work.

10.21437/SMM.2018-3

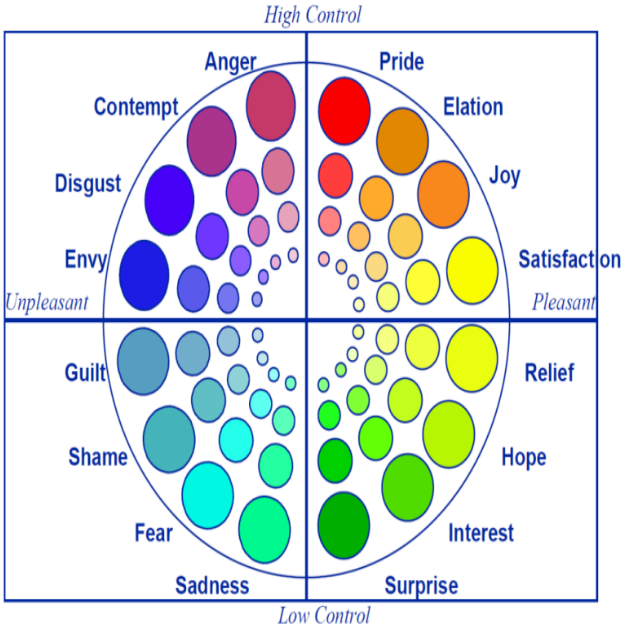

Figure 1: *Geneva emotion wheel depicting different emotion classes [17].*

2. Realistic Database and Emotion Annotation Process

This section describes the details of real-world data used in this work and the annotation process considered for marking speech emotions. The annotation process with emotion priority and confidence level is inspired from GEW structure. The segmentation process used for the continuous speech segmentation is also discussed here.

2.1. Realistic Speech Corpus

In this work, a conversational speech data collected in real-world scenario has been used for the studies. This corpus includes day to day conversation from a large population over telephone channel in English language. Approximately 10 hours of data collected in such condition is used for the analysis. The data contains 29 different telephonic conversations. Automatic segmentation is carried out on each conversation. A total of around 13000 segments are obtained which are then used for emotion labeling and analysis. A team of 10 members having scientific background with basic speech knowledge are considered as labelers for the study. Due to confidentiality involved with this database, we are unable to share the database publicly at this moment.

2.2. Geneva Emotion Wheel: Different Speech Emotions

The GEW is a theoretically derived and empirically tested instrument to measure emotional reactions to objects, situations and events [16]. It consists of different emotion classes over a wheel structure that are aligned in a symmetric manner. Sixteen out of twenty distinct emotions are used in this work. Figure 1 shows the layout of a Geneva wheel, where 16 different classes are depicted over a circle. It has two axes that denote the different emotion behavior. The horizontal axis is for valence, which indicates attractiveness or pleasantness of an event. Negative valence corresponds to unpleasant and positive valence corresponds to pleasant emotions. The vertical axis refers to arousal or control, where level of arousal increases as we move from bottom to top of the wheel. For example, the emotion anger happens due to high arousal as well as it is an unpleasant state. Therefore, its position in the Geneva Wheel is in the fourth quadrant. The emotions having similar aspects are in the same quadrant of the circle which can be seen from the figure. There are five levels confidence shown in terms of small circular structure for each emotion in the GEW that shows the amount of confidence level for a particular emotion.

Algorithm 1 : Segmentation algorithm

```
1: Compute short term energy of speech using window size of
   20 ms with a shift of 20 ms.
2: Segment onsets and offsets are marked at the instants where
   the signal energy crosses 0.02 times the average signal energy.
3: Adjacent segments are merged if the gap between the boundaries
   are less than 150 ms. (This is performed to remove the spurious
   gaps produced due to the low energy sounds such as stops, weak
   fricatives etc.)
4: If the duration of a segment is less than 400 ms and if it is
   at least 700 ms away from the adjacent segments, then the
   segment is discarded. (This removes the isolated spurious
   segments of very short duration.)
5: If the duration of a segment is less than 2 s and it's adjacent
   segment is present within 700 ms, then those adjacent segments
   are merged to make a single segment. (This ensures a minimum
   segment duration of 2 s for most of the segments.)
6: Segments are saved ensuring at least 150 ms silences region
   before and after the detected segment onsets and offsets.
```

2.3. Automatic Speech Segmentation

It is very difficult to mark a single emotion label for a very long speech file. The signal may contain different emotions at different instants of time. Therefore, it is essential to split a longer speech signal into smaller segments. The segmentation is performed based on signal energy. Short term energy is computed and segment boundaries are marked at the instants where the signal energy crosses 0.02 times of the average signal energy. Subsequently, some segments are merged and some are removed depending on the size of the segments and the gap between the segments. The detailed segmentation procedure is shown in Algorithm 1.

Fig. 2 illustrates the segmentation process. Fig. 2 (a) shows portion of a speech signal and the segments (in dark rectangles) obtained by thresholding the short term energy of the signal. Spurious gaps which are less than 150 ms are removed by merging the adjacent segments. The spurious segments which have less than 400 ms duration and are occurring at least 700 ms away from the adjacent segments are discarded. The refined segments are shown in Figure 2 (b) using the dotted rectangles along with the speech signal. Segments are further refined by merging the adjacent segments which are not more than 700 ms away and are of less than 2 s in duration. Refined segments are shown in Figure 2 (b) using the dark rectangles. This further refinement ensures duration of most of the segments greater than 2 s. After obtaining the segments, the segmentation boundaries are saved in a label file (.lab) for processing in Wavesurfer to mark emotion classes.

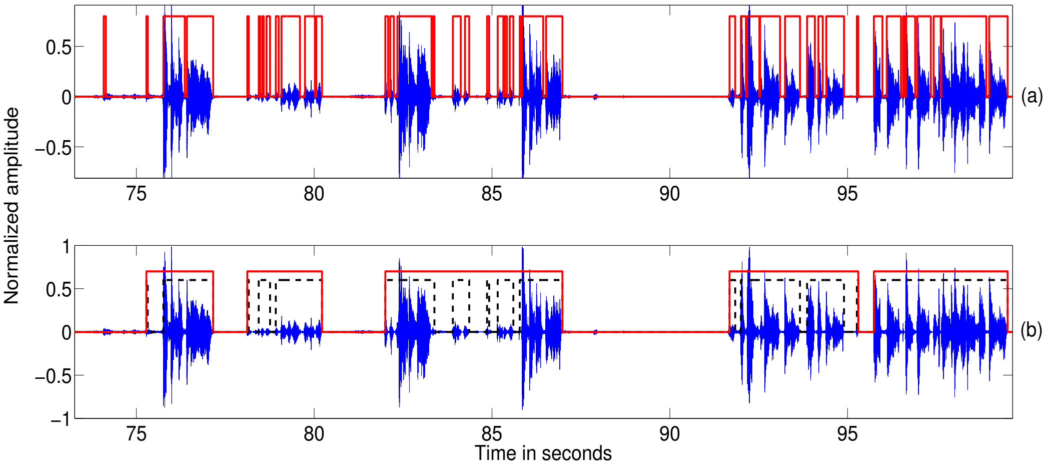

Figure 2: *Figure illustrating segmentation process. (a) A portion of a speech signal along with the segments shown in dark rectangles obtained using short term energy. (b) Speech signal and the segments after refinement.*

Table 1: *Statistics of different emotion classes on realistic data annotated using customized labelling.*

Emotions	#Occurrence (%)	Avg. Confidence Scores
Neutral	62.29	4.9
Pride	0.70	4.6
Elation	0.85	4.8
Joy	4.59	4.7
Satisfaction	1.04	4.8
Relief	0.47	4.4
Hope	0.92	4.5
Interest	1.88	4.2
Surprise	1.34	4.5
Sadness	4.82	4.7
Fear	0.30	4.5
Shame	0.16	4.3
Guilt	0.47	4.7
Envy	0.24	4.9
Disgust	1.98	4.1
Contempt	4.70	4.6
Anger	13.24	4.7

2.4. Annotation

The annotation is performed using Wavesurfer tool as mentioned earlier. For each segment obtained using the automatic segmentation method as explained in previous subsection, the user marks the emotion class for the particular segment based on perceptual knowledge. As perception of emotion is highly subjective and varies from one user to another, we have proposed a customized annotation process for labelling. This customized process involves emotion priority and a confidence score. The emotional state of a person are very loosely differentiated as we go around the GEW. Therefore, the user is required to provide three emotion labels to a particular segment on a priority order. Further, a confidence score in the range 1 to 5 has to be given for the three priority emotion class on another panel in the Wavesurfer. In this way, combining the priority emotion and confidence level score we intend to derive a more appropriate emotion label for a particular segment of data collected in real-world environment. It is also to be noted that if the user finds that a particular segment has no specific emotion, then one can mark it as neutral speech.

3. Studies and Discussion

In this section, we make a study on the realistic data used for emotion analysis. It is important to know the significance of different emotions present in day to day conversational speech. This can help in choosing relevant emotion classes in a practical setting. Further, finding the confusions among different classes is also useful as that can help in differentiating one class from another along with possibilities of evolving different methodologies for discriminating confusing emotion categories. The studies in this work primarily focus along these two directions.

Table 1 shows the statistics of different emotions on the annotated realistic data used in this work. Percentage of occurrence of the emotions which are labeled with the highest confidence score are shown in the table. It can be observed that most of the speech in day to day conversation is neutral (around 62%). Only 38% of speech are marked as one of the 16 emotions shown in Figure 1. Among these emotions, joy, anger, sadness and contempt are observed to be marked frequently compared to other classes. Average confidence scores corresponding to the first priority based emotion class are also shown in the third column of Table 1. It is observed that confidence level of the labelers are on the higher side while marking the neutral class. Further, it is observed that confidence level is lower in case of emotion classes disgust, interest and shame.

Table 2 shows the confusion among different emotion classes during annotation. The confusion matrix is computed by calculating the probability of second priority emotion class given the first priority emotion class. It can be seen that almost all emotion classes are confused with the neutral class. On the other hand, once the neutral is marked as the first priority, it is less likely that any other emotion class is marked as the second priority. Among different emotion classes, anger, joy and disgust, when marked as the first priority class, are confused

Table 2: *Confusion matrix showing percentages of second priority emotion classes given the first priority emotion class during annotation. Boldfaces are showing the diagonal entries and red colors are showing the major confusions.*

	Neutral	Pride	Elation	Joy	Satisfaction	Relief	Hope	Interest	Surprise	Sadness	Fear	Shame	Guilt	Envy	Disgust	Contempt	Anger
Neutral	**91.1**	0.2	0.2	1.4	0.2	0.2	0.3	0.5	0.3	2.7	0.1	0.0	0.2	0.0	0.3	0.6	1.8
Pride	17.6	**42.6**	2.9	1.5	5.9	0.0	1.5	4.4	0.0	0.0	0.0	0.0	1.5	2.9	8.8	1.5	8.8
Elation	16.9	5.1	**16.9**	13.6	18.6	1.7	10.2	8.5	1.7	3.4	0.0	0.0	0.0	3.4	0.0	0.0	0.0
Joy	7.9	1.0	6.4	**70.0**	4.6	1.5	1.0	3.1	1.3	0.5	0.0	0.0	0.5	0.0	0.3	1.8	0.3
Satisfaction	28.6	5.5	3.3	5.5	**25.3**	4.4	12.1	2.2	0.0	7.7	0.0	0.0	0.0	2.2	2.2	0.0	1.1
Relief	14.6	0.0	0.0	0.0	22.0	**39.0**	4.9	7.3	0.0	12.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hope	21.8	0.0	1.3	2.6	2.6	3.8	**41.0**	14.1	1.3	9.0	0.0	0.0	0.0	1.3	0.0	1.3	0.0
Interest	20.3	1.4	4.7	6.1	2.0	0.0	9.5	**34.5**	6.1	8.1	0.0	0.0	0.0	2.7	0.7	2.0	2.0
Surprise	16.4	0.0	2.6	0.9	3.4	0.9	0.9	8.6	**47.4**	1.7	2.6	0.9	0.9	2.6	5.2	2.6	2.6
Sadness	26.5	0.5	0.3	0.0	0.3	0.5	3.2	1.6	1.1	**30.2**	3.2	1.9	9.0	0.5	3.7	6.4	11.1
Fear	14.3	0.0	0.0	3.6	0.0	0.0	0.0	0.0	3.6	14.3	**32.1**	10.7	10.7	0.0	10.7	0.0	0.0
Shame	7.1	0.0	0.0	0.0	14.3	0.0	0.0	0.0	0.0	14.3	0.0	**50.0**	7.1	0.0	7.1	0.0	0.0
Guilt	7.9	0.0	2.6	0.0	0.0	0.0	0.0	0.0	2.6	23.7	0.0	5.3	**42.1**	0.0	0.0	10.5	5.3
Envy	19.0	4.8	4.8	4.8	0.0	0.0	14.3	14.3	0.0	0.0	4.8	0.0	0.0	**14.3**	9.5	4.8	4.8
Disgust	2.0	1.3	0.0	0.0	0.3	0.0	0.3	0.0	2.3	2.0	0.0	2.0	0.7	0.7	**81.6**	2.6	4.3
Contempt	25.8	0.6	0.6	0.0	0.0	0.0	0.0	1.8	1.2	14.7	0.6	1.2	1.2	1.2	16.0	**31.3**	3.7
Anger	4.0	0.8	0.5	0.2	0.1	0.0	0.0	0.0	0.4	9.8	0.1	0.2	0.4	2.5	11.9	0.4	**68.8**

lesser compared to the other classes. Further, it can be seen that elation is confused with joy and satisfaction, relief is confused with satisfaction, hope is confused with interest, fear is confused with sadness, shame and guilt, shame and guilt are confused with sadness, and contempt is confused with disgust. Except the confusions with neutral, most of the other confusions are occurring near the diagonal of the confusion matrix. This is expected because the emotions around the diagonal are placed closer in the emotion wheel.

The reported studies in this work are few preliminary investigation of speech emotion on real-world data. These analyses can form a basis towards future explorations on emotion detection in realistic environment. Further, the customized annotation process having emotion priority and confidence level can be useful while dealing with unlabeled data collected in practical settings. The future work will focus on developing emotion recognition systems for day to day conversations using this annotated speech corpus and its comparison to acted emotion based systems.

4. Conclusion

This work discusses the need for investigating emotions on realistic environment data from the view of potential applications. The importance of annotating emotions in real-world condition is highlighted as the emotion classes are very confusing in such scenarios. In this regard, a customized annotation process using Wavesurfer tool is proposed that involves emotion priority and confidence level. The emotion priority involves top three emotions identified by the user and the confidence level provides a score in a scale of 5 to depict the associated confidence. Further, an automatic segmentation method is proposed to segment the long conversational speech into smaller segments for annotation process. The studies are made on a database collected on a day to day conversational speech over telephone channel. In this work, analysis on the occurrences of different emotion classes on real-world scenario has been investigated. Additionally, the emotions well distinguishable to other classes and the most confusing ones are studied that may provide some insights for future explorations to handle the sensitive cases. The future work will focus on extending this work to build emotion recognition systems using the annotated practical data and then to evaluate on conversational speech in realistic environments.

5. Acknowledgement

The authors would like to thank the members and the interns of BAAAI, Guwahati for carefully annotating the realistic data used in this work. The work of the second author is supported by a project grant in NUS with WBS no: R-263-000-D22-597. The work of remaining authors is supported by Kaliber Labs.

6. References

[1] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[2] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2227–2231.

[3] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204, 2016.

[4] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 223–227.

[5] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Communication*, vol. 53, no. 9, pp. 1062 – 1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing.

[6] M. E. Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572 – 587, 2011.

[7] M. Neumann and N. T. Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” *CoRR*, vol. abs/1706.00612, 2017. [Online]. Available: http://arxiv.org/abs/1706.00612

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *INTERSPEECH*, 2005.

[9] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 12 2008.

[10] F. Chenchah and Z. Lachiri, “Speech emotion recognition in noisy environment,” in *2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) 2016*, March 2016, pp. 788–792.

[11] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, “Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks,” in *Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association*, San Francsico, CA, September 2016, pp. 3593–3597.

[12] A. Tawari and M. M. Trivedi, “Speech emotion analysis: Exploring the role of context,” *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502–509, Oct 2010.

[13] www.speech.kth.se/wavesurfer/, “Wavesurfer,” -.

[14] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, “Introducing the recola multimodal corpus of remote collaborative and affective interactions,” in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE), in Proc. of IEEE Face and Gestures 2013, Shanghai (China)*, 2013.

[15] http://www.speech.kth.se/hat/, “Higgins annotation tool,” -.

[16] K. Scherer, V. Shuman, J. Fontaine, and C. Soriano, “The grid meets the wheel: assessing emotional feeling via self-report,” in *Components of emotional meaning : a sourcebook*, ser. Series in affective science, J. Fontaine, K. Scherer, and C. Soriano, Eds. Oxford University Press, 2013, pp. 281–298.

[17] K. R. Scherer, “What are emotions? and how can they be measured?” *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.