Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Literature Review and
Statement of Research Intent

Project 26

Emotional speech

visualization and

annotation

Author: Sunny Choi

Project Partner: Enuri Kolugala

Supervisor: Jesin James

Co-supervisor: Felix Marattukalam

28/04/2023

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Name: Hye Seon (Sunny) Choi

ABSTRACT: With the rapid advancements of contactless devices and robot automation fields, speech is prevalently utilized as a means of interaction between computer systems and humans in diverse areas of technology. There are many researchers in the field of speech technology who focus on the recognition of emotions in speech signals (SER) to enable computer systems to respond more effectively to humans. One of the major issues encountered in SER is the lack of annotation tools available for speech emotion which leads to insufficient training data required for the development of deep learning algorithms for SER. Thus, this project aims to develop a web-based speech emotion annotation and visualisation tool that is capable of creating a speech emotion database in an efficient manner. A literature review is conducted to gain an understanding of fundamental knowledge on emotion in speech signals, highlight challenges encountered with emotion annotation and visualisation, evaluate different existing annotation tools and suggest potential areas that could be refined on EmotionGUI.

1. Literature Review

This section elaborates on extensive research areas of interest that are identified and evaluated to acquire an overview of the current state of Speech Emotion annotation and visualisation tool development.

1.1 Nature of emotion in the speech signal

The human brain serves as the locus of multiple emotions simultaneously, albeit with one dominant emotion expressed at any given moment [1]. These blended emotions may vary consistently over time and sometimes exhibit an abrupt shift in their emotional state from one to another. Due to the fusion of multiple emotions, it is often said that it is inherently subtle and cannot be defined as a single emotional state. Emotional data can be captured via multiple modes, such as speech, gestures, facial expressions and biosignals [1]. This project focuses on emotional content conveyed via audio or visual stimulus.

1.2 Application of emotion in HCI

All interactions among humans are primarily associated with emotions, which denotes that the approach to the development of HCI should consider emotion as a core aspect. Emotion embedded into the computer system helps obtain additional contextual information about the situation where the system is in and enables the system to interact and correspond with a user in a much more natural and pleasant manner, that is more human-like [2].

1.3 Emotion models

Both categorical and dimensional models provide different approaches to representing emotions in speech signals.

1.3.1 Discrete emotion models

With categorical emotional models, one emotion is selected out of a set of emotions that indicates the best feeling experienced by an annotator [3]. Emotions can be labelled by making use of Ekman's six basic emotions or Plutchik's eight core emotions [1]. The model with Ekman's six emotions has been adopted by most research in affective computing [3]. Limitations regarding emotion classification have been clearly presented as categorical models are unable to encompass all the emotions into defined sets of categories [1]. Furthermore, there may be cases where identical emotional states fall into different emotional categories because of personality, cultural or environmental variations etc., which make it challenging to find the actual category where it belongs [4]. Therefore, these limitations denote that emotional categories may not provide distinct sets of different emotional states, which further brings out errors in emotion detection [3]. Another issue has been noted in a case where there is no appropriate class to select from the label set. Regardless of the presence of these limitations, categorical models and their variations have been employed prevalently due to their easily understandable emotion labels and casualness.

1.3.2 Dimensional emotion model

On the other hand, dimensional models are employed to represent emotional states based on a set of quantitative measures using multidimensional scaling [3]. Different sets of dimensions, where each dimension represents a distinct feature of emotions, constitute various emotional states in the model. A two-dimensional (2D) model, as the most widely used dimensional model, represents levels of Valence (unpleasant-pleasant) and Arousal (sleepiness/ boredom-excitement) independently [1]. One of the most well-known 2D models is Russell's circumplex model, which incorporates 2D polar coordinates of Valence-Arousal (V-A) levels organized in a circumplex shape [4]. Common applications of three-dimensional (3D) models are Mehrabian's model using PAD and Self-Assessment Manikin (SAM). SAM is a picture-oriented approach that depicts levels of three dimensions with aids of facial images, and thus the image aids, considered as a language-free measure, define the emotions as easily understandable and make the model widely applicable as being used across various settings with different populations, gender, age, race, and cultural backgrounds [5]. Continuous dimensional models, in general, are a favourable approach to measuring distinct emotional states as they are capable of capturing a wide range of fine emotion concepts, providing a means for estimating the similarity between emotional states [3]. Also, like the feature of blended emotion in real life, emotional states in dimensional space are related to each other, which is hugely different from categorical models [3].

1.4 Challenges of emotion annotation

Annotation of emotions in speech signals is a significantly challenging task caused by various factors in different aspects. Firstly, a factor contributing to the difficulties can be elaborated with respect to the subjectivity of annotation tasks [6]. It is common for humans to hold differing opinions, which results in seeing the same subject from different perspectives to others. Since annotation tasks are strongly dependent on how annotators interpret emotional contents perceived in verbal data, it is less likely to expect consistent emotion labelling between annotators [7]. According to Bayerl [8], the inter-annotator agreement score can be improved by providing annotators with more training so that they gain familiarity with using the annotation tool. A contradictory opinion is presented by Mohammad [8] that over-training would have an adverse effect on the agreement score as it led to more confusion and apprehension in judgement tasks. The nature of emotion being dynamically and constantly changing impedes the annotation process to some extent, as it is typically challenging to capture the rise and fall of emotion that varies over time [1]. Types of data that are being annotated also impact the level of difficulty of emotion annotation. Out of three types of corpora used; acted, induced and natural, most speech emotion databases are built based on acted data, with their emotional content often seen as exaggerated, conveying a single emotion [9]. This achieves a better inter-annotator agreement score than those of other types but makes the dataset less valuable for real-world application due to the nature of mixed emotions in real life [9]. It is ideal for working with a natural database, which reflects closer to realistic behaviours; however, only a small portion of relatively weak emotions are apparent in natural data, compared to distinct emotions of acted speech [1]. Further studies are required to apply the natural database in the speech emotion sector.

1.5 Review of existing emotion annotation and visualisation tools

Comparisons among existing emotion annotation and visualisation tools are made by those listed in Table 1, in chronological order, providing information about their features and availability of analysis review. Modified Transcriber [1] is excluded from the comparison as the annotation scheme provided, such as sentence-by-sentence marking on five different emotional states, is considered heavily time-consuming and tedious in comparison to the other tools. AffectRank is also not included in the comparison due to very poor statistical analysis results, found in section 1.6.2, which make the tool unreliable in practical use.

Table 1: Existing Emotion Annotation Tools with features(from left to right) including published year, annotation type (continuous or frame by frame), emotional model (ratings on one- or two-dimensions, or category), customization options (customizable emotional scale or other features), annotation tool (mouse, joysticks or keyboard), operating systems, whether code is open-sourced, local installation requirement or web-based, and visualization ratings

Tools	**Year**	**Annotation Type**	**Emotion Model**	**Custom**	**Annotation Tool**	**Platform**	**Open-sourced?**	**Installation required?**	**View ratings**
FEELTrace [10]	2000	Continuous	2D	-	Mouse	Windows XP	No	Yes	-
Modified Transcriber [1]	2005	Sentence by sentence/ MECAS	Multi-categories, 1D	-	Mouse	All	No	Yes	-
EMuJoy [11]	2007	Continuous	2D	✓	Joystick, mouse	Java	Yes	Yes	-
GTrace [12]	2012	Continuous	1D	✓	Mouse	Windows	Yes	Yes	-
ANNEMO [13]	2013	Continuous	1D	-	Mouse	All	No	Web-based	-
CARMA [14]	2014	Continuous	1D	✓	Mouse, keyboard	Windows	Yes	Yes	✓
AffectRank [15]	2015	Continuous	2D	-	Mouse	-	No	Not available	-
VA Online Annotation Tool (VAOAT) [7]	2017	Frame by frame	1D	-	Mouse, keyboard	All	No	Web-based	-
DANTE [16]	2017	Continuous	1D /AffectButton	-	Mouse	All	No	Web-based	-
JERI [17]	2017	Continuous	2D/SAM	-	Joystick	-	No	Not available	-
DARMA [18]	2017	Continuous	2D	✓	Joystick	Windows	Yes	Yes	✓
EmotionGUI	2022	Continuous	2D	-	Mouse	All	Yes	Yes	✓

NB: A complete version of this table with extensive features included is available via the GitHub project link which is provided upon request.

1.6.1 Comparisons for existing emotion annotation and visualisation tools

As shown in Table 1, most annotation tools are continuous dimensional models, except for VAOAT, which marks emotional states frame by frame. The per-frame annotation leads to a bias creation, resulting in sharp annotation signals that conflict with the dynamic expression of emotion [19]. Thus, it is believed by many researchers that continuous annotation is better at describing the simultaneous dynamic nature of emotions. All the tools listed in Table 1 adopt two dimensions of valence and arousal, yet there is always debate about measurements to be made on one dimension (1D) at a time or two dimensions (2D) simultaneously. The 1D model is employed to alleviate the cognitive load caused by recordings in 2D bipolar space [1]. On the other hand, a key advantage of 2D measurements claimed is that the 2D framework allows for much richer emotional descriptions than those provided by 1D alone, by providing coverage of intermediate states defined by blends of two dimensions [20]. FEELTrace [10] was the first tool to address issues associated with gradation and variation of emotional states over time. Approaches to deal with the issues were implemented by introducing a colour-coded cursor derived from Plutchik’s emotional index [20], which offers an intuitive way for users to associate with the relevant emotional state. Representation of time progression was applied by gradually shrinking the circle pointer indicating the current mouse position in the space over time. These features were also adopted by GTrace and EmotionGUI. GTrace introduced a special feature that allows users to work on emotional scales other than valence and activation by providing various sets of emotional terms to be selected by users [12]. Users are now able to collect measurements of any kind to adapt to nearly any project dealing with continuous ratings [20]. This feature is also evident in EMuJoy, Carma and Darma. ANNEMO, VAOAT and DANTE started to introduce a web-based version to facilitate remote annotation, allowing users to log into their accounts in the web interface [13]. The web interface enables a faster annotation process and easier access to the tool without needing to undergo all the processes required for installation [12]. There are tools adopting a pictorial representation of emotional scale in their interfaces, which were DANTE using AffectButton and both EMuJoy and JEEI using SAM [8]. This provided users with more intuitive labelling and prior training was no longer required [12]. Carma and Darma added a window for viewing previously collected measurements, and Darma, in particular, had an additional feature, which offered various rating analysis options for estimating descriptive statistical results. These features made significant improvements in the efficiency, training, and quality control of research tasks [14]. Darma, with all its essential features such as playback-measurements auto-synchronization, annotation review tools and easy customization, makes itself surpass all existing 2D systems.

1.6.2 Validation of the annotation tools

Annotation tools involved in a systematic analysis include FEELTrace, DANTE, ANNEMO, AffectRank and DARMA. Statistics resulting from different tools are not comparable to each other because the analysis for each tool was conducted using different statistical metrics under different settings such as database, participator, and media files to be rated etc. Thus, the interpretation of statistics is made independently in the aspect of reliability or usability depending on the methods of analysis carried out. Table 2 presents statistical analysis data from different tools, of which each interpretation is explained as follows:

- FEELTrace: Statistical analysis of annotations was carried out using ANOVA[10]. If the F value is much bigger than 1 and the P value is less than 0.05, the difference among groups is deemed statistically significant [10]. Three comparisons were performed considering different sets of groups: the first one between neutral and emotional starts, the second one between low activation and high activation and the last between low evaluation and high evaluation. All the results clearly state that the differences among groups are highly significant. Thus, these results indicate that FEELTrace is capable of distinguishing neutrality and each of the four emotional states very reliably. However, [10] reported that it failed to differentiate between fear and anger.

- DANTE: ICC of emotional dimensions estimate the agreement between annotators, and those obtained values shown in Table 2 imply that, on average, all the annotators agreed on their ratings to a great extent. Cronbach's α, Corr. and CCC were also performed to assess inter-annotator agreement. The results yielded low values for arousal and high values for valence; which means ratings for valence are reliable, whereas arousal was more difficult to distinguish, resulting in a poorer agreement between raters. C_α is used to estimate the consistency between annotations; $\alpha>0.7$ is considered an acceptable consistency, and $\alpha>0.8$ as a good consistency [13]. Thus, C_α of 0.842 for valence is considered a good internal consistency, whereas that of arousal is 0.325. As a result, DANTE can be regarded less reliable to differentiate various levels of arousal.
- ANNEMO: C_α results for valence (0.74) and arousal (0.80) show that the consistency between annotations is good for arousal and acceptable for valence. Mean correlation values for both dimensions were found within the range between 0.4 and 0.59 [13], so they are both considered to have moderate associations. These results denote that ANNEMO can be used with high reliability for distinguishing between dimensions of arousal and valence.
- AffectRank: In [15] conducted analysis comparing rating-based FEELTrace and raking-based AfterRank, and Table 2 includes analysis results from [15] only if provided with an exact number. As Cronbach's α was only applicable to the unidimensional model, it only applied to FEELTrace, and the result indicated good intra-agreement reliability between annotations. Krippendorff's α analysis was carried out for both FEELTrace and AfterRank to estimate the degree of inter-agreement between annotators. The results of **K_α**^{b} for AfterRank show very low consistency between annotators for both dimensions, and those for FEELTrace indicated even lower agreement. Given such low **K_α**^{b} values resulted for AffectRank, this tool is considered unreliable for differentiating between valence and arousal.
- DARMA: In [18], a reliability testing of auto-synchronization was carried out by measuring the time delay between consecutive joystick samples. The analysis indicated good reliability of the synchronization, supported by results stating that 80% of samples fall in the range of target delay ± 0.001s [18]. Evaluation of user satisfaction regarding DARMA v5.00 or later versions was also conducted through an online survey using a seven-point scale. Responses to the survey indicated users' significant satisfaction with DARMA across all the aspects of software, such as annotation, reviewing ratings, and website, except for software installation and

documentation, given slight to moderate satisfaction. Reviewing the responses from the survey, DARMA achieved usability to a great extent.

Table 2. Statistics of annotations from different tools using various metrics; ANOVA, intraclass correlation coefficient (ICC), Mean correlation coefficient (Corr.), Concordance correlation coefficient (CCC), Cronbach's α (C_α), Krippendorff's α (K_α)

FEELTrace [10]			**DANTE** [16]					**ANNEMO** [13]		**AffectRank** [15]	
ANOVA	**F(23,1)**	**P**		**ICC**	**C_α**	**Corr.**	**CCC**	**Corr.**	**C_α**	**C_α**[a]	**K_α**[b]
Emotions - Neutral	453	<0.001	Pleasure (Valence)	0.664	0.842	0.742	0.509	0.407	0.74	0.9	0.33
Low-High Activation	239	<0.001	Arousal	0.921	0.325	0.310	0.093	0.435	0.80	0.69	0.16
Low-High Evaluation	847	<0.001	Dominance	0.948							

[a]: Cronbach's α value from FEELTRACE

[b]: Krippendorff's α value from AffectRank

1.6.3 Trends in current studies on emotion annotation tool

Many researchers are primarily interested in continuous dimensional annotation tools [20], which is clearly reflected in Table 1. Reviewing all the features in Table 1 and supported by responses to the survey conducted in [18], there is no preference observed for annotation tools used, or whether development code is open-sourced. However, there is a tendency shown towards making a transition to a web-based application allowing easier accessibility and better usability. Tools developed in later years are not limited to Windows but are compatible with all operating systems, including Windows, Linux, and Mac. Many studies recognize the benefit of having customizable emotional dimensions and have started to employ it in their tools a decade ago. In addition, having the capability of making annotations by multiple annotators is essential now. The review ratings window added to DARMA seems very beneficial for training data in SER.

1.6 Possible improvement on EmotionGUI

In response to the survey conducted on EmotionGUI, further improvement should focus on clarity of representations of multiple annotations and emotional gradations by colour change. Other updates are planned to implement the tool on the website, and, possibly, expand a customizable option for emotional dimensions by incorporating 1D frameworks. UI can also be modified more fitted to web-based version.

2. Research Intent

2.1 Project scope

The main focus of our research will be to develop a web-based speech emotion annotation and visualisation tool to aid in the creation of a speech emotion database that will contribute to training data in SER. The following questions will guide our research :

- How can we design a web-based system that can represent variations and gradation of emotional states over time effectively?
- What are the features highly recognized and acknowledged by many studies in the development of speech emotion annotation and visualisation tools these days?

2.2 Project objectivities

To answer the questions listed above, we have identified steps to be taken according to various focus areas of the tool.

- Literature review:
  - Conduct a review of existing speech emotion annotation and visualisation tools
  - Conduct research on evaluation techniques for emotion annotation and visualisation tools
  - Explore a currently existing emotion annotation and visualisation tool called 'EmotionGUI'
- Annotation
  - Explore ways to provide users with a better representation of changes in emotional states with respect to time
- Visualisation
  - Explore ways to allow users to utilize various emotion recognition models and visualise predicted emotional data on a 2D plane, presented on a single webpage.
- Live audio recording
  - Explore ways to represent speech signals of recorded audio data graphically on the user interface

2.3 Project plan

As EmotionGUI is a desktop GUI application programmed using PyQt, my background knowledge and experience in GUI application development using PyQt will facilitate the implementation of tasks related to the annotation and live audio recording parts by implementing existing PyQt codes in a web-based setting. On the other hand, my partner, Enuri, has knowledge of UX and HCI concepts and experience with the application of HTML, CSS and JavaScript in web design. Thus, she is responsible for the tasks associated with visualization and overall UX consideration to be made in our system.

3. Conclusions

Through a literature review, knowledge and relevant content on speech emotion were covered in detail. It elaborated on various challenges confronted by annotating emotional states and evaluated existing emotion annotation and visualisation tools, from which several suggestions on potential areas were made to improve EmotionGUI across different aspects. From reviewing various existing tools, it was noted that different features are selected to be applied to a tool according to the tasks required to be achieved and implemented.

Acknowledgements

I would like to thank the supervisor Jesin James and the co-supervisor Felix Marattukalam for providing the support and guidelines to make this project successful.

References

[1] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. 2005 Special Issue: Challenges in real-life emotion annotation and machine learning based detection. Neural Netw. 18, 4 (May 2005), 407–422. https://doi.org/10.1016/j.neunet.2005.03.007

[2] Lichtenstein, A., Oehme, A., Kupschick, S., Jürgensohn, T. (2008). Comparing Two Emotion Models for Deriving Affective States from Physiological Data. In: Peter, C., Beale, R. (eds) Affect and Emotion in Human-Computer Interaction. Lecture Notes in Computer Science, vol 4868. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-85099-1_4

[3] S. PS and G. Mahalakshmi, “Emotion models: a review,” International Journal of Control Theory and Applications, vol. 10, no. 8, pp. 651–657, 2017.

[4] Beale, R., Peter, C. (2008). The Role of Affect and Emotion in HCI. In: Peter, C., Beale, R. (eds) Affect and Emotion in Human-Computer Interaction. Lecture Notes in Computer Science, vol 4868. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-85099-1_1

[5] Bynion, T. M., & Feldner, M. T. (2017). Self-Assessment Manikin. Available: https://link.springer.com/content/pdf/10.1007/978-3-319-24612-3_77.pdf

[6] Öhman, E. (2020). Emotion Annotation: Rethinking Emotion Categorization. *DHN Post-Proceedings*. Available: https://www.semanticscholar.org/paper/Emotion-Annotation%3A-Rethinking-Emotion%C3%96hman/72c953db90d98c7377a942394161aa127ba1494d>

[7] Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image and vision computing*, *65*, 23-36. https://doi.org/10.1016/j.imavis.2017.02.001

[8] Bynion, T. M., & Feldner, M. T. (2017). Self-Assessment Manikin. Available: https://scholar.google.com/scholar?cluster=2641290485391308982&hl=en&as_sdt=0,5#d=gs_cit&t=1679363167597&u=%2Fscholar%3Fq%3Dinfo%3ARwRZNU34zrUJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D1%26scfhb%3D1%26hl%3Den

[9] Vogt, T., André, E., Wagner, J. (2008). Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation. In: Peter, C., Beale, R. (eds) Affect and Emotion in Human-Computer Interaction. Lecture Notes in Computer Science, vol 4868. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-85099-1_7

[10] Cowie, Roddy & Douglas-Cowie, E. & Savvidou, Suzie & McMahon, E. & Sawey, M. & Schr\"oder, M.. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. Proceedings of the ISCA Workshop on Speech and Emotion.

[11] Nagel, F., Kopiez, R., Grewe, O., & Altenmüller, E. (2007). EMuJoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, *39*(2), 283-290.

[12] Cowie.R, Sawey.M. (2013): GTrace. School of Phychology, Queen’s University, Belfast, United Kingdom

[13] F. Ringeval, A. Sonderegger, J. Sauer and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, 2013, pp. 1-8, doi: 10.1109/FG.2013.6553805.

[14] Girard, Jeffrey. (2014). CARMA: Software for Continuous Affect Rating and Media Annotation. Journal of Open Research Software. 2. e5. 10.5334/jors.ar. Available: https://www.researchgate.net/publication/262936715_CARMA_Software_for_Continuous_Affect_Rating_and_Media_Annotation

[15] G. N. Yannakakis and H. P. Martínez, "Grounding truth via ordinal annotation," *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, 2015, pp. 574-580, doi: 10.1109/ACII.2015.7344627.

[16] Boccignone, G., Conte, D., Cuculo, V., & Lanzarotti, R. (2017, November). AMHUSE: a multimodal dataset for HUmour SEnsing. In *Proceedings of the 19th ACM international conference on multimodal interaction* (pp. 438-445).

[17] K. Sharma, C. Castellini, F. Stulp and E. L. van den Broek, "Continuous, Real-Time Emotion Annotation: A Novel Joystick-Based Analysis Framework," in *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 78-84, 1 Jan.-March 2020, doi: 10.1109/TAFFC.2017.2772882.

[18] Girard, J.M., C. Wright, A.G. DARMA: Software for dual axis rating and media annotation. *Behav Res* **50**, 902–909 (2018). https://doi.org/10.3758/s13428-017-0915-5

[19] Ramakrishnan, S., El Emary, I.M.M. Speech emotion recognition approaches in human computer interaction. *Telecommun Syst* **52**, 1467–1478 (2013). https://doi.org/10.1007/s11235-011-9624-z

[20] Wang, Yan & Song, Wei & Tao, Wei & Liotta, Antonio & Yang, Dawei & Li, Xinlei & Gao, Shuyong & Sun, Yixuan & Ge, Weifeng & Zhang, Wei & Zhang, Wenqiang. (2022). A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances.