Time to Embrace Emotion Change: Selecting Emotionally Salient Segments for Speech-based Emotion Prediction

Zhaocheng Huang[1,2] and Julien Epps[1,2]
[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2] Data61, CSIRO, Australia
zhaocheng.huang@unsw.edu.au, j.epps@unsw.edu.au

Abstract

Continuous prediction of emotion dimensions has gained popularity recently, because systems of this kind can capture subtle changes in emotions in naturalistic settings. However, most of these systems take utterance-level or frame-level features as input, without considering within-utterance variation in emotion. This paper investigates data selection for speech-based emotion prediction from an emotion change perspective, and finds that annotation delays vary even within utterances. Experimental results on the RECOLA corpus show that emotion-change frames carry relatively greater emotion-related information, achieving 5.4% and 24.5% relative improvements over the baseline for arousal and valence prediction under the Output-Associative Relevance Vector Machine Framework.

Index Terms: data selection, annotation delays, emotion changes, continuous emotion prediction, relevance vector machine

1. Introduction

Within the affective computing community, there has been a trend towards representing emotions in terms of arousal and valence dimensions, which are numerical values representing how activated a person is and how pleasant they feel [1]. They are considered a more descriptive representation of complex and subtle emotions in naturalistic environment, compared with conventional emotion categories, such as neutral and anger [2]. This trend has been further driven by the annual Audio/Visual Emotion Challenges (AVEC) [3], which targets continuous prediction of arousal and valence.

Continuous emotion prediction is a regression problem, which involves feature extraction and regression modelling. More specifically, features are extracted from training utterances and then used to train a regression model, based on which features extracted from testing data can be used to generate predictions. Among possible regression models, the Support Vector Regression (SVR) and Relevance Vector Machine (RVM) have been shown to be successful for this task [4]. In this work, RVM is preferred because an advanced framework based on RVM, called Output-Associative (OA) RVM, has shown promise on AVEC data [5]. The OA-RVM comprises two stages of RVM regression modeling. The first RVM is trained on input features. Temporal arousal and valence predictions from the first RVM are then associated with input features for RVM training at the second stage.

During regression modelling, it is well-recognized that there exist annotation delays in emotion ratings, which are manually assigned by annotators. The delay is mainly caused by reaction lag between annotators' perceptual observations and decision-making [6], as well as fatigue or variations in attention. This delay has a great impact on emotion prediction system performance, and correct annotation delay compensation has been associated with quite dramatic improvements in accuracy [4].

Although there have been extensive investigations into emotion recognition during last decade, most studies neglect within-utterance variation and treat all parts equally. This may not be a good assumption in general, and improvement in performances of speech-based emotion prediction systems requires a better understanding of emotionally-salient segments within speech. Attempts to investigate this previously include examining emotion recognition accuracies using specific phonemes or phoneme classes, where it was found in [7], [8] that vowels, especially /a/ are more conducive to emotion classification, whilst Bitouk et al. [9] suggested that spectral features extracted from consonants are more effective. Le et al. [10] examined various data selection strategies based on classifier agreement, and compared utterance selection with sub-utterance selection (segments within utterances) in terms of emotion classification performances, as well as convergence rate and stability in training process.

Kim [11] speculated that another way to identify emotionally salient segments is to explore variations in emotions: low variation in emotions implies clear expression of emotions, which may favor emotion recognition. However, this have not been experimentally investigated, and recent studies suggest that short-term emotion dynamics can facilitate emotion classification [12] and diagnosis of psychological diseases [13]. In continuously annotated emotional corpora, it is observed that emotion ratings tend not to change across time, which leads to a large proportion of frames without changes in emotion ratings. Motivated by [12], [13], as well as this latter observation, it is reasonable to pose the question: are frames with emotion change more salient for predicting emotions?

2. Database

The database used in this paper is Remote Collaborative and Affective Interactions (RECOLA), a spontaneous multimodal corpus collected in settings where two French speakers remotely collaborate to complete a survival task via a video conference. During the collaborative interactions, multimodal signals, including audio, video and physiological signals such as ECG and EDA, were collected from 46 participants (data from 23 participants are publically available). This database is chosen because it is a large, high quality database that has been continuously annotated at every 40 milliseconds for arousal and valence by six annotators. Moreover, the recent

AV+EC2015 challenge [3] employed a subset of the database (18 speakers in total and 5-minute recording per speaker), which was evenly partitioned into training and development sets for a continuous emotion prediction task. In this study, we considered only speech signals and the same partitions as used in AV+EC 2015.

3. Data Selection

3.1. Defining Partitions based Emotion Change

This section defines different database partitions based on emotion changes, i.e. changes in the arousal and valence ground truth provided in the RECOLA corpus. To separate the "change" frames from "non-change" frames, first-order differences from the emotion ratings were calculated, as seen in Fig.1, where all data are partitioned into three parts: B ("before"), C ("change"), and A ("after"). This data partition scheme is applied to arousal and valence separately.

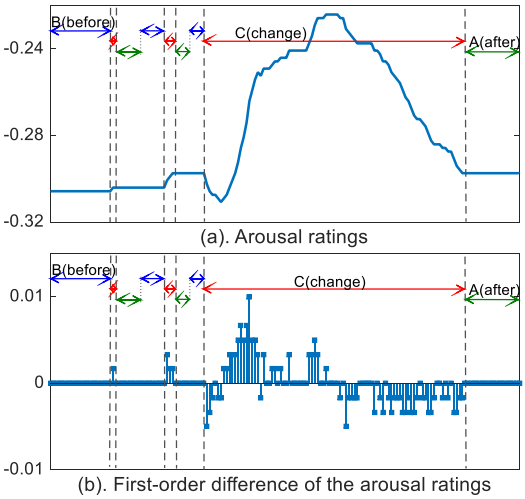

Figure 1: *All data are divided into B, C, and A based on the first-order difference where zeros mean no emotion change, whilst non-zeros mean emotion change. B and A were separated in order to understand differences between before and after emotion changes. These is a large proportion of "non-change" frames.*

As shown in Fig.1, partition C contains all the frames where emotion ratings change in addition to all frames whose ratings remain unchanged for less than L frames. Partition B contains all the frames before emotion ratings change at the beginning of every file, and the second half of all frames whose ratings remain unchanged for more than L frames. Similarly, partition A contains all the frames after the last change frame of every file, and the first half of all frames whose ratings remain unchanged for more than L frames. L is the minimum number of frames considered for non-change frames, i.e. B and A. This parameter is introduced to provide more continuity for C, as seen in Figure 1. With L = 1 frame, partitions B, C and A account for 16.86%, 63.18%, and 19.96% respectively for arousal. For valence, partitions B, C, and A account for 19.36%, 58.10%, and 22.54% respectively.

3.2. System Overview

In addition to feature extraction and regression modelling, the proposed system includes data selection and delay compensation. Data selection partitions all training data into three subsets: B, C and A. Delays are compensated via fixed temporal shifts and smoothing at the training and testing phase respectively, as per [4]. The regression model used in this paper is RVM, chosen because it offers good performance with fast training time. More importantly, it has shown promise across various system settings within the OA-RVM framework.

For training the RVM, we employed the *SparseBayes MATLAB toolbox*, and the only parameter to be tuned is the iteration number, selected from between 10 and 30. The temporal window size for construction of output-associative matrices containing input features and spatial temporal predictions was fixed to 151 frames as in [4]. The features used were 88-dimensional eGeMAPS functionals [3], extracted using a 2 second window size every 40 milliseconds, to align with the emotion ratings. All features were scaled into [0, 1] before training, and scaling coefficients from training data were used to normalize testing data. To ensure comparability, data selection was only conducted on training data, and all results are reported using all test data. During delay compensation, optimum delay values were chosen from [0, 6] s with 0.4 s increments. Performances of emotion prediction systems were measured using Concordance Correlation Coefficients (CCC) [3].

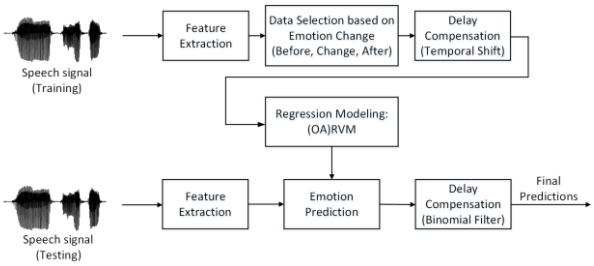

Figure 2: *Proposed system for investigating data selection based on emotion change.*

3.3. Emotion Prediction using Subsets of Training data

This section compares performances of different systems trained using either all training data or different subsets from Section 3.1, namely B, C, or A, all with a global delay compensated, tested on all test data (with no partitioning). The global delays, estimated from different delay values trialled on all training data and test data, as seen in Figure 2, were found to be 3.2s for arousal and 3.6s for valence.

Table 1: *Comparisons of performances using either all data (B+C+A) or subsets of training data (B, C, or A).*

	B+C+A	B	C	A
Arousal	0.60	0.19	0.52	0.20
Valence	0.33	0.09	0.31	0.11

As shown in Table 1, the system trained on only change frames (C only) performed comparably to the system using all data (B+C+A), especially for valence. With the caveat that partition C contains a relatively larger amount of data than B or A, this suggests that frames where emotion ratings change carry more emotion-related information that favors emotion prediction.

Since annotation delay is important, and may vary between different partitions, it is perhaps unwise to keep a global delay value for different training partitions. This motivates us to search for the optimum delays for different training partitions (i.e. B, C and A).

3.4. Annotation Delay Optimized for Data selection

This experiment investigates how different delay values impact system performances when a subset of training data is used for training. The global delays estimated in section 3.2 were fixed for all test data (which remained unchanged throughout), whereas optimum delays for selected training partitions were trialled from among [0, 6] s with 0.4 s increments arousal and valence respectively. This approach was adopted throughout the following experiments.

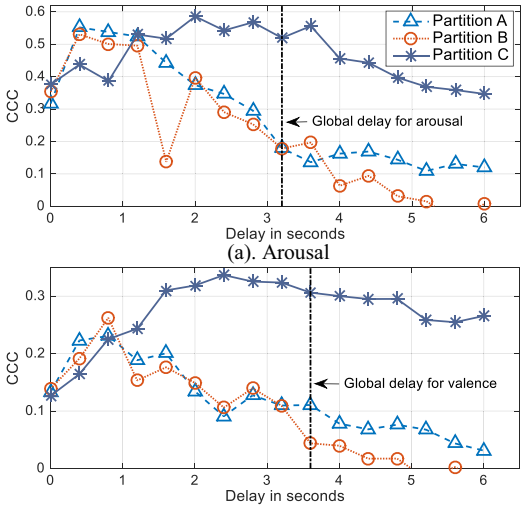

Figure 3: *Delay compensation for different data partitions (A, B and C), selected based on the first-order difference of arousal and valence ratings.*

As seen in Figure 3, the optimum annotation delays (for this database) were found to be around 0.4-0.8s for partitions B or A and around 2-3s for C, which were consistent for arousal and valence. This suggests that annotation delays are different not only among various annotators [6], but also different across time within-annotator. The greater delays for C are because of annotators taking more time to react to emotion change. Moreover, optimum delays for partition C are close to the global delays, suggesting that the delays mainly come from people's reactions to emotion change (C), not B or A, which is expected but, to the best of our knowledge, has not been shown in literature before. It also suggests that the global delays are influenced more strongly by change speech segments than non-change speech segments.

Systems trained on C only with optimum delay provided equivalent performances to systems trained on all data in Section 3.3. We repeated the above experiments using the OA-RVM framework, comparing systems trained on B+C+A with those trained on C only: 0.71 vs 0.71 for arousal, 0.40 vs 0.41 for valence. Since the OA-RVM framework consistently provided better performances than RVM (used in Fig. 3), the OA-RVM was used throughout all experiments below.

4. Emotion Change based Data Selection

4.1. Smoothed deltas vs First-order difference

In the above systems, we partitioned the training data based on the first-order differences of emotion ratings. However, the first-order difference is problematic because: (i) it assumes the possibility of extremely rapid emotion changes (i.e. the sampling interval between ratings is 0.04 seconds), which are unrealistic; (ii) raters tend not to move their cursor continuously, which results in a very large proportion of zero values, as shown in Fig.1(b); (iii) there is annotation noise caused by annotator tremble [14]. To resolve this, we proposed smoothing the first-order differences using a Moving Average (MA) filter, followed by applying a threshold to select "large emotion changes". The window size of the MA filter, herein referred to as W (measured in frames), needs to be chosen: the larger W, the smoother the changes in emotion ratings. In order to find the best W, we compared different values from 10 to 240 frames for arousal and valence respectively, as seen below in Fig.4.

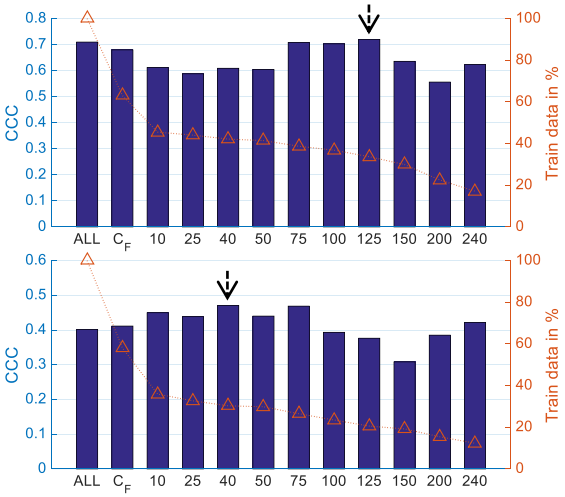

Figure 4: *"ALL" means baseline system trained on B+C+A, whereas "C_F" means a baseline system trained on only C (selected from the first-order difference). Other systems used C only based on smoothed deltas with different W values. The arrows indicate W values chosen to arousal and valence.*

A benefit offered by smoothed differences is that the change value for each single frame is a collective decision from all the frames within the smoothing window rather than only a change between two adjacent frames. This can reduce annotation noise, and a large value potentially suggests that the majority of frames within the temporal window are changing. To this end, smoothed differences are more beneficial than first-order differences for selecting emotion change. As shown in Figure 4, compared with first-order differences (C_F), smoothed differences provide better performances. The best W values for arousal and valence were 125 frames (CCC=0.72) and 40 frames (CCC=0.47) respectively, which were retained through the following experiments. This may suggest that arousal prediction is favored within a large region where emotion ratings in the majority of frames are changing, whereas valence prediction is more effective within smaller regions (10 - 75 frames).

Moreover, notice that although C_F, change frames based on first order difference, accounts for 63.18% and 55.10% of all the training data for arousal and valence respectively, systems with around 30-40% of the training data provided the best performances. This may suggest that large changes are more informative for emotion prediction, which motivates investigation into large emotion changes.

4.2. Large Emotion Changes

This section investigates emotion prediction systems trained on large emotion change frames. To do this, we applied a

thresholds T to select large changes, which however leads to a reduction in C-partition training data. To mitigate this problem, we included adjacent frames around the large emotion changes for training. This leads to different combinations of Ts and number of adjacent frames around the large emotion changes, as seen in Figure 5.

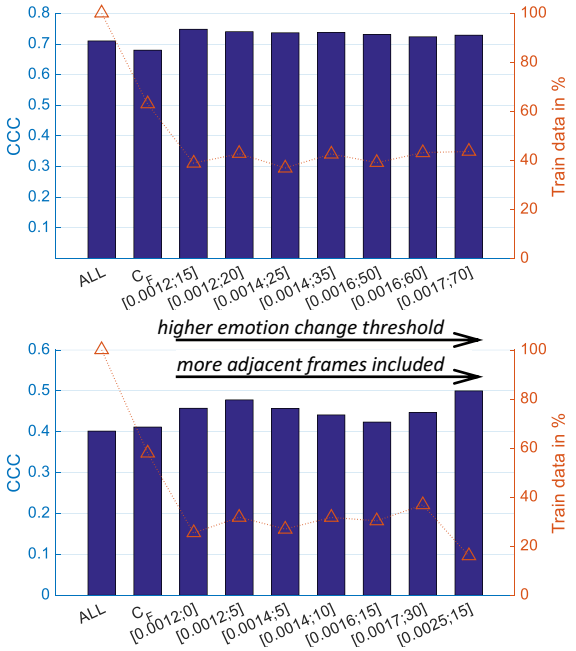

Figure 5: *Performances of systems trained on large emotion changes, compared with two baselines "ALL" and "C_F". [0.0012, 5] means selecting all emotion changes larger than T = 0.0012 and considering the nearest 5 frames adjacent to large changes. Percentages of resultant partitions with only large changes were shown in the orange markers.*

In Figure 5, we gradually increased thresholds to eliminate frames with small variations (which leads to less data) and included adjacent frames to maintain sufficient data for training (around 20% - 50% of total training data). It can be seen from Fig. 5 that systems trained on regions where large emotion changes occur yielded better performances over the baseline trained on all training data, achieving 0.75 vs 0.71 for arousal and 0.50 vs 0.40 for valence. Notice that this is achieved using 38.95% and 16.06% of training data for arousal and valence, lending support to the hypothesis that large emotion changes are more salient for emotion prediction.

Furthermore, in order to check how well the results generalise, the best settings for arousal and valence were tested via 6-fold cross-validation (15 speakers for training and 3 speakers for testing per fold). For arousal, W=125 frames, with [0.0012, 15] for large changes and 2.8 s delay for selected training data at each fold. For valence, W=40 frames, with [0.0012, 5] and 2.4 s delay for selected training data. System performances under these settings were 0.72 for arousal, and 0.41 for valence. The performances do not generalise very well. This is presumably due to the fixed delays, which are better to be optimized in training data within each fold.

5. Conclusions and Future Work

This paper has investigated data selection based on emotion changes for speech-based emotion prediction systems. Experimental results consistently show that speech segments containing emotion changes are more salient for emotion prediction (especially for valence). Training on only Change (C) frames gives comparable performances for arousal prediction and slightly better performances for valence, compared with performances using all data. When large emotion change (C) frames were used for training, after smoothing first-order differences, we achieved 5.4% and 24.5% improvements in CCC for arousal and valence relative to the baseline. This is significant because valence prediction from speech is generally recognized to be a difficult problem in the literature [15].

Moreover, this paper experimentally demonstrates that delays in emotion perception, reflected in annotation, mainly arise from people's reactions to emotion change (C), not non-change segments (B or A).

This paper is limited in that only one database has been tested. Future work involves extending this investigation, i.e. data selection based on emotion change, to multiple databases.

6. Acknowledgement

This work is partly funded by Data61, CSIRO.

7. References

[1] Gunes, H. and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vis. Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[2] Cowie, R., G. McKeown, et al., "Tracing Emotion," *Int. J. Synth. Emot.*, vol. 3, no. 1, pp. 1–17, 2012.

[3] Ringeval, F., B. Schuller, et al., "AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in 5th *AV+EC, ACM MultiMedia*, 2015.

[4] Huang, Z., T. Dang, et al., "An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction," in *AVEC*'15.

[5] Nicolaou, M. a., H. Gunes, et al., "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image Vis. Comput.*, vol. 30, no. 3, pp. 186–196, 2012.

[6] Marioryad, S. and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, 2014.

[7] Lee, C., S. Yildirim, et al., "Emotion recognition based on phoneme classes.," in *INTERSPEECH*, 2004, pp. 889–892.

[8] Sethu, V., E. Ambikairajah, et al., "Phonetic and speaker variations in automatic emotion classification," in *INTERSPEECH*, 2008, pp. 617–620.

[9] Bitouk, D., R. Verma, et al., "Class-level spectral features for emotion recognition," *Speech Commun.*, vol. 52, pp. 613–625, 2010.

[10] Le, D. and E. Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *ACII*, 2015.

[11] Kim, Y., "Exploring sources of variation in human behavioral data: Towards automatic audio-visual emotion recognition," *Affect. Comput. Intell. Interact. (ACII),* 2015.

[12] Provost, E., "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *ICASSP*, 2013.

[13] Houben, M., W. Van Den Noortgate, et al., "The Relation Between Short-Term Emotion Dynamics and Psychological Well-Being: A Meta-Analysis.," *Psychol. Bull.*, vol. 141, no. 4, pp. 901–930, 2015.

[14] Metallinou, A. and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," *in 10th International Conference on FG* 2013.

[15] Grimm, M., K. Kroschel, et al., "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, pp. 787–800, 2007