INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Parametric Distributions to Model Numerical Emotion Labels

Deboshree Bose, Vidhyasaharan Sethu, Eliathamby Ambikairajah

School of Electrical Engineering and Telecommunications
University of New South Wales, Sydney
NSW 2052, Australia

deboshree.bose@unsw.edu.au,v.sethu@unsw.edu.au,ambi@ee.unsw.edu.au

Abstract

It is common to represent emotional states as values on a set of numerical scales corresponding to attributes such as arousal and valence. Often these labels are obtained from multiple annotators who record their perception of emotion in terms of these attributes. Combining these multiple annotations by taking the mean, as is typical in affective computing systems ignores the inherent ambiguity in the labels. Recently it has been recognised that this ambiguity carries useful information and systems that employ distributions over the numerical scales to represent emotional states have been proposed. In this paper we show that the common and widespread assumption that this distribution is Gaussian may not be suitable since the underlying numerical scales are bounded. We then compare a range of well-known distributions defined on bounded domains to ascertain which of them would be the most suitable alternative. Statistical measures are proposed to enable quantifiable comparisons and the results are reported. All comparisons reported in the paper were carried out on the RECOLA dataset.

Index Terms: emotion ambiguity, continuous emotion prediction, inter-rater variability, probabilistic modeling

1. Introduction

Emotions are complex and their expression is often subtle [1]. Consequently, their perception can be ambiguous at times [2, 3, 4]. An obvious manifestation of this ambiguity is that emotions are perceived somewhat differently by the individual observers depending on their individual history, mood, personality, culture, their involvement, context, the environment and physiological factors [5]. Emotion databases typically consist of multimodal (speech, video, EEG, etc.) recordings of emotionally coloured expressions or interactions along with one or more set of emotion labels associated with the recording as perceived by the raters. Most widely used datasets include labels obtained from multiple annotators and the ambiguity inherent in the perception of emotions is evident in the varying levels of disagreement between these ratings. Emotion labels may be categorical (e.g., 'Angry', 'Happy', etc.), values on numerical scales (e.g., arousal/activation - calm vs. excited, valence - negative vs. positive) [6, 7].

Most affective computing systems employ some method to combine the multiple labels into a single label, for e.g., by taking the majority label when dealing with categorical labels or taking the mean rating when dealing with numerical arousal/valence ratings [8, 6]. The reasoning behind this is that a 'mean' rating is a better representation of the perceived emotion as the perceived emotion is evenly felt across all raters. Most nuanced approaches include estimating inter-rater agreement to discard annotations that deviate most from the mean of other annotations before taking the weighted average annotation as the best estimate for true emotion [9]. When combining multiple labels this way, the individual variations in emotion ratings are considered noise and penalized. However, concepts of annotation noise, averaged annotations or dominant emotions are at best reductionist assumptions that may be very unrealistic in many practical instances. Emotion representations need to be capable of reflecting the diversity of human annotation, due to the inherently subjective nature of affective experiences, both while expressing and perceiving emotions [10].

There are some examples of emotion representations in affective computing that incorporates this diversity as ambiguity in perceived emotions. Likewise the focus of this paper, these typically involve modelling the set of ratings with a suitable distribution in the space of numerical arousal-valence ratings. Most commonly, Gaussian distributions are employed for their computational convenience [11, 12, 13], but other alternatives include Gaussian mixture models [14, 15], Gaussian processes [16, 17], and other non-parametric models [18]. However, despite the computational convenience, assuming that numerical emotion ratings are distributed normally may not be accurate since the numerical scales themselves are bounded (for e.g., valence ranging from -1 to 1) [13] and a symmetrical distribution might not be the best fit for the ratings. In this paper, we define and employ quantitative measures to ascertain the appropriateness of assumption of Gaussianity that is often made. Furthermore, we compare a range of parametric distributions defined on a bounded domain to the Gaussian distribution and with each other.

2. Modeling Ambiguity with Gaussians

As previously mentioned, the arousal and valence scales along which emotions are rated numerically are all bounded, while the support of a Gaussian distribution is over all real numbers. Consequently, representing the distribution of a set of emotion ratings as a Gaussian implicitly implies that there is a non-zero probability that emotion rating can be outside the allowable range. In practise, this might be an acceptable approximation as long as the variance of the distribution is small, but it is also worth noting that the Gaussianity assumption is less likely to be suitable when the numerical labels are close to the edge of the interval of allowable values. This is illustrated in Figure 1 which shows an example each for arousal (blue) and valence (red) where the Gaussian fit to the six ratings also implies there is a non-trivial probability of the rating falling outside the interval $[0, 1]$, which does not reflect reality (ratings are bounded to be within $[0, 1]$).

To quantify the extent of this potential issue, we analyse the RECOLA dataset which includes a set of continuous time-varying arousal and valence ratings obtained from six independent raters, sampled every $40ms$, as emotion labels [4]. In the RECOLA dataset both arousal and valence ratings are con-

http://dx.doi.org/10.21437/Interspeech.2021-1000

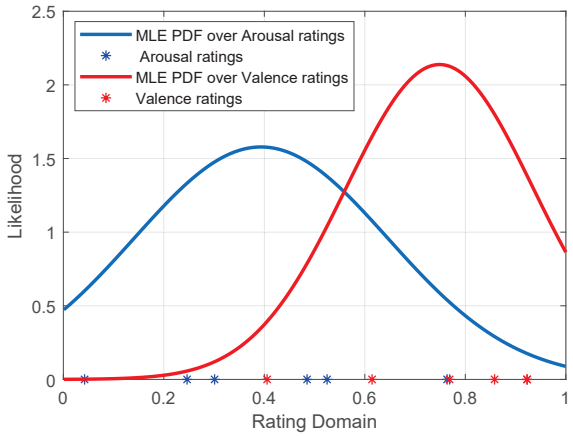

Figure 1: *Examples frame from RECOLA showing six arousal and valence ratings as well as maximum likelihood Gaussian distribution fit to the ratings. In both cases, the best fitting distributions do not agree with the fact that the ratings are also bounded.*

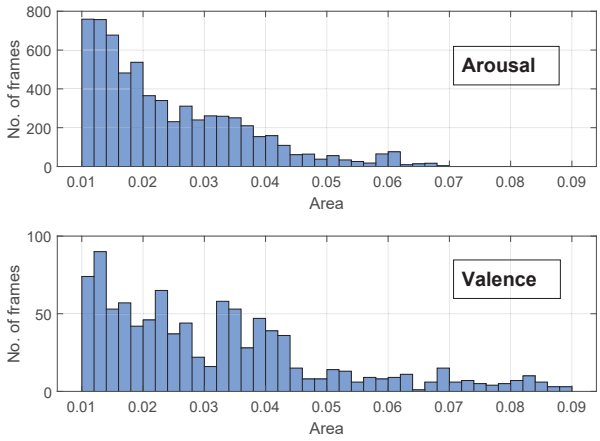

Figure 2: *Histograms constructed from all instances or frames of emotion ratings when at least by 1 percent of the total area under the Gaussian Maximum Likelihood distributions lie outside the domain* $[0, 1]$.

strained to the interval $x \in [-1, 1]$. For convenience (distributions with bounded support, refer section 3, are usually described by assuming the interval of support is $[0, 1]$), and without any loss of generality, we apply a linear transform ($y = 0.5x + 0.5$) to map the ratings to the interval $y \in [0, 1]$. Following this, for every frame ($40ms$) we estimate the maximum likelihood Gaussian fit to the six arousal (and valence) ratings,

$$\boldsymbol{\theta_{ML}} = \arg\max_{\theta=[\mu,\sigma]} \prod_k \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mathbf{y}_k-\mu)^2}{2\sigma^2}} \tag{1}$$

where, y_k denotes the k^{th} arousal/valence rating, $k = 1, ..., 6$, and $\boldsymbol{\theta}$ denotes the mean (μ) and standard deviation (σ) of the Gaussian.

Next, we compute the probability implied by this Gaussian distribution that the rating falls outside $[0, 1]$, $P_{y\notin[0,1]}$, as:

$$P_{y\notin[0,1]} = 1 - \int_0^1 \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{-\frac{(y-\hat{\mu})^2}{2\hat{\sigma}^2}} dy \tag{2}$$

where, $\hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood mean and standard deviation estimates for the Gaussian, that is, $\boldsymbol{\theta_{ML}} = [\hat{\mu}, \hat{\sigma}]$.

An analysis on the RECOLA dataset revealed that for almost 10% of data, the maximum likelihood (ML) Gaussian fit to arousal ratings indicates a greater than 1% chance that arousal ratings will fall outside the allowed interval ($[0, 1]$). Similarly, the ML Gaussian fits to valence ratings for more than 1.5% of the data indicates a greater than 1% chance that valence rating will fall outside $[0, 1]$. Figure 2 shows a histogram of the $P_{y\notin[0,1]}$ for all the instances where $P_{y\notin[0,1]} > 0.01$. Finally, it is worth noting that this measure does not indicate goodness of fit within $[0, 1]$, which is explored in section 3.

3. Comparing Distributions

The analyses in section 2 suggests that a distribution with bounded support may be more appropriate than a Gaussian to model a distribution over numerical emotion ratings on bounded scales. In this section, we examine an extensive list of commonly used parametric distributions with bounded support (refer Table 1), having reasonable shapes and degrees of freedom that would allow for the ambiguity in the perceived emotion ratings to be captured. These distributions are compared to each other and to the Gaussian based on how well they fit the ratings from each annotator which is quantified in terms of log-likelihoods. These comparisons are carried out using the best fitting distribution parameters, obtained as maximum likelihood estimates (MLE), as well expected value over the entire parameter posteriors. Note that most of the distributions listed in Table 1 have two parameters except the Trapezoidal distribution which has four parameters and the Logit Metalog family of distributions which admits a variable number of parameters with a greater degree of freedom in its shape with increase in number of parameters. We evaluate the Logit Metalog with 2 and 3 parameters.

3.1. Comparing Maximum Likelihood Estimates

Intuitively the most straightforward approach towards comparing these distributions is to determine the MLE of the distribution parameters at each point and compute the overall log-likelihood across the entire dataset. We estimate the maximum log-likehood for each distribution, $\mathcal{L}_{ML}$, as:

$$\mathcal{L}_{ML} = \frac{1}{N} \sum_n \sum_k \ln p(y_{n,k}|\boldsymbol{\psi}_{n,ML}) \tag{3}$$

where, $y_{n,k}$ denotes the k^{th} annotator's rating at time frame n; N is the total number of time points in the database; and $\boldsymbol{\psi}_{n,ML}$ is MLE estimate of the distribution parameters at n:

$$\boldsymbol{\psi}_{n,ML} = \arg\max_{\psi} \prod_k p(y_{n,k}|\boldsymbol{\psi}) \tag{4}$$

The average maximum log-likelihoods estimated for all the distributions of the RECOLA dataset for arousal, $\mathcal{L}_{\mathcal{A}}$, and valence, $\mathcal{L}_{\mathcal{V}}$, are listed in the first two columns in Table 1.

Additionally, we make sure that the overall log-likelihood score over the entire database is not being skewed by a few extreme data points; we also plot the histogram of the frame-wise log-likelihood ratios between two distributions to compare them. We expected the histogram to reveal a preponderance of

positive values when the first distributions fits better than the second for a preponderance of data points.

3.2. Expected Log-Likelihood

A challenge with comparing distributions based on the MLE of the distribution parameters, as outlined in section 3.1, is that it may be skewed by overfitting. This is of particular significance in this problem since the number of annotators, and consequently ratings, is usually very small (the RECOLA dataset only has 6 ratings) and the risk of overfitting cannot be overlooked. That is, there may be certain parameters $\boldsymbol{\psi}$ where the log-likelihood of the ratings given those parameters may be uncharacteristically high whereas other similar parameter values would be associated with a low log-likelihood. Furthermore, the chance of overfitting increases with distributions with more parameters, such as the Trapezoidal distribution. To overcome this limitation, we estimate the expected log-likelihood (ELL) as follows:

$$E_{p(\boldsymbol{\psi}|\mathbf{y}_n)}[ln(p(\mathbf{y}_n|\boldsymbol{\psi}))] = \int_{\psi} p(\boldsymbol{\psi}|\mathbf{y}_n) \ln(p(\mathbf{y}_n|\boldsymbol{\psi}))\, d\boldsymbol{\psi} \quad (5)$$

where, $E_{p(\boldsymbol{\psi}|\mathbf{y}_n)}[\cdot]$ denotes the expected value with respect to the posterior probability over the distribution parameters, $p(\boldsymbol{\psi}|\mathbf{y}_n)$.

Further, we can see that by applying Bayes' theorem and rearranging the terms in (5), we get

$$\begin{aligned} E_{p(\boldsymbol{\psi}|\mathbf{y}_n)}[ln(p(\mathbf{y}_n|\boldsymbol{\psi}))] = &\int_{\psi} p(\boldsymbol{\psi}|\mathbf{y}_n) \ln(p(\boldsymbol{\psi}|\mathbf{y}_n))\, d\boldsymbol{\psi} \\ &- \int_{\psi} p(\boldsymbol{\psi}|\mathbf{y}_n) \ln(p(\boldsymbol{\psi})\, d\boldsymbol{\psi} \\ &+ \int_{\psi} p(\boldsymbol{\psi}|\mathbf{y}_n) \ln(p(\mathbf{y}_n))\, d\boldsymbol{\psi}. \end{aligned} \quad (6)$$

Since, for any ζ, $\int_{\zeta} p(\zeta|\mathbf{y}_n)\, d\zeta = 1$, the third term in (6) simplifies to $ln(p(\mathbf{y}_n))$.

$$\begin{aligned} E_{p(\boldsymbol{\psi}|\mathbf{y}_n)}[ln(p(\mathbf{y}_n|\boldsymbol{\psi}))] = &-H[p(\boldsymbol{\psi}|\mathbf{y}_n)] + ln(p(\mathbf{y}_n)) \\ &- \int_{\psi} p(\boldsymbol{\psi}|\mathbf{y}_n) \ln(p(\boldsymbol{\psi})\, d\boldsymbol{\psi} \end{aligned} \quad (7)$$

Integrating the third term by parts we have,

$$\int_{\psi} p(\boldsymbol{\psi}|\mathbf{y}_n) \ln(p(\boldsymbol{\psi}))\, d\boldsymbol{\psi} = ln(p(\boldsymbol{\psi}) - \int_{\psi} \frac{p'(\boldsymbol{\psi})}{p(\boldsymbol{\psi})}\, d\boldsymbol{\psi} \quad (8)$$

Substituting the result of (8) in (7),

$$E_{p(\boldsymbol{\psi}|\mathbf{y}_n)}[ln(p(\mathbf{y}_n|\boldsymbol{\psi}))] = -H[p(\boldsymbol{\psi}|\mathbf{y}_n)] + ln(p(\mathbf{y}_n)) + C \quad (9)$$

where, C is a constant of integration.

From (9), we see that the proposed measure $E_{p(\boldsymbol{\psi}|\mathbf{y}_n)}[p(\boldsymbol{\psi}|\mathbf{y}_n)]$ depends upon the entropy H of the posterior distribution. A distribution with lower H, is sharper and since ELL depends on $-H$, a higher ELL indicates a better fit over a range of parameter values.

4. Experimental Settings

4.1. RECOLA Database

The Recola multimodal database [4] consists of over 9.5 hours of spontaneous dyadic conversation recordings in French of which, the expressed affective states in the first 5 minutes of each conversation is annotated by six French-speaking assistants and by the participants themselves. Of these, 18 conversations were provided as part of the Audio-Visual Emotion Recognition Challenge (AV+EC 2016) [19] with 9 distinct utterances in the training and development partitions each. Amongst the most widely used and publicly available datasets employed in continuous emotion prediction research, the RECOLA dataset has the largest number of individual annotations per sample (six annotations). Consequently, we chose it for all the analyses reported in this paper. All 18 utterances from the training and development partitions of the AV+EC 2016 [19] were used in this study. 6 individual time continuous annotations sampled at 40ms was provided for arousal and valence for each utterance. Each rating is a numerical value in the interval $[-1, 1]$. As outlined in section 2, we map these to the interval $[0, 1]$ for convenience.

Table 1: *Aggregate measures over all frames*

Distribution	$\mathcal{L}_A$	$\mathcal{L}_V$	E_A	E_V
Gaussian	6.0329	8.1728	5.0340	7.1299
Beta	6.1208	8.2510	**5.1190**	**7.2804**
Truncated Gaussian	6.0548	8.1778	5.0336	7.0984
Logit Normal	6.0774	8.2345	4.6924	6.6195
Kumaraswamy	6.1160	7.9443	5.1044	7.0481
Raised Cosine	6.0658	8.1464	4.2919	6.4628
Trapezoidal	7.6986	10.1690	4.7127	6.7080
Logit Metalog 2 param.	5.9902	8.1277	4.7860	6.7603
Logit Metalog 3 param.	5.9931	8.1180	4.7497	6.6831

4.2. Prior Choices

Distribution parameters could be classified as: location parameters describing the distribution's position in its region of support, and scale and/or shape parameter(s) describing the distribution's spread and shape. Location parameters were assigned uniform priors since the region of support is bounded and there is no reason to prefer one position over another.

Inverse-gamma priors were chosen for scale parameters as in hierarchical models [20] for every Gaussian-like distribution because Inverse-Gamma prior has a desirable closed form expression due its conditional conjugacy [21] versus Half-Cauchy densities, and the uniform distribution which were implicitly assumed for the MLE estimates. Gamma priors were chosen for shape parameters as in [20].

In the experiments described in this paper, all relevant quantities of interest for all distributions were computed numerically over a uniform grid over the parameter space. Additionally, the leading frames of each utterance, where the ratings were all zero valued, were discarded.

5. Results and Discussion

The average log-likelihood based on maximum likelihood fits for the various distributions, $\mathcal{L}_{\mathcal{A}}$ and $\mathcal{L}_{\mathcal{V}}$, and the expected log-likelihood over the parameter posteriors, E_A and E_V, are listed in Table 1. In terms of the log-likelihood of the MLE fits, the Trapezoidal distribution best describes the ratings, followed by the Beta distribution represented as, $\boldsymbol{\beta}$. However, it is worth remembering that the Trapezoidal distribution has four parameters (significantly greater than for the other distributions here), and the ML estimate is obtained using only 6 ratings and con-

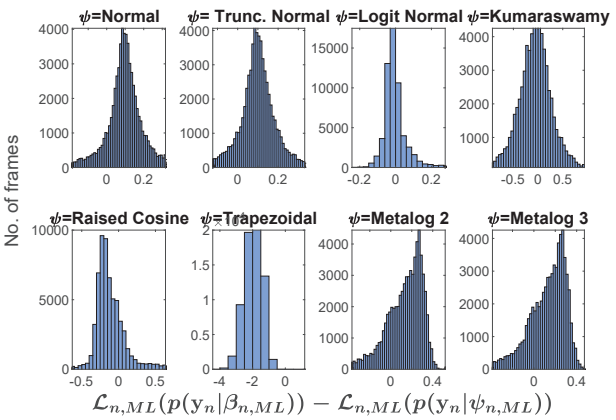

Figure 3: *Histograms of log-likelihood ratios of Beta distributions to others based on ML parameter estimates for arousal ratings.*

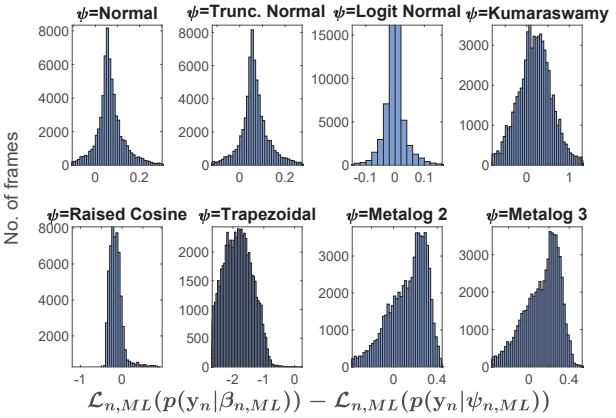

Figure 4: *Histograms of log-likelihood ratios of Beta distributions to others based on ML parameter estimates for valence ratings.*

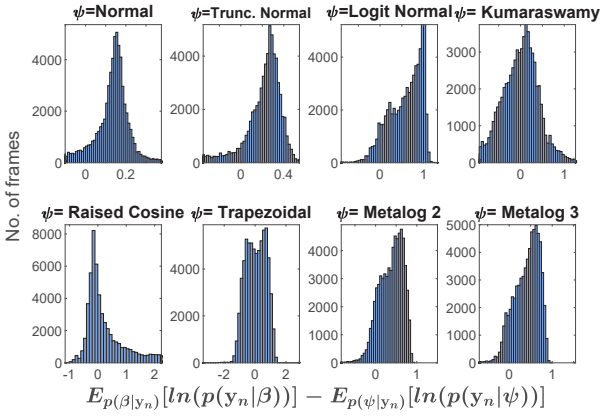

Figure 5: *Histograms of expected log-likelihood ratios of Beta distributions to others over parameter posteriors for arousal ratings.*

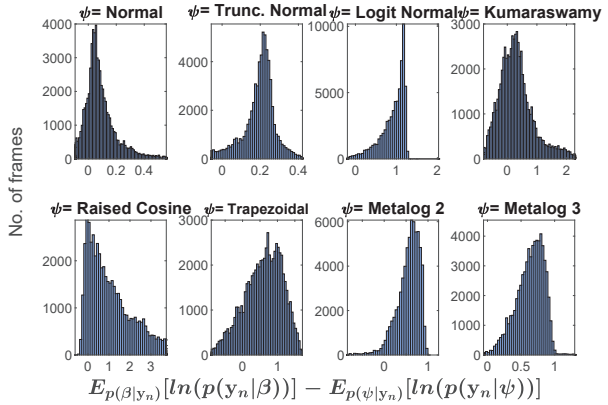

Figure 6: *Histograms of expected log-likelihood ratios of Beta distributions to others over parameter posteriors for valence ratings.*

sequently there is a strong chance of overfitting. Looking at the expected log-likelihood values, we can see this indeed might be the case. Overall based on Table 1 it would be reasonable to suggest the Beta distribution (and the closely related Kumaraswamy distribution) is the best choice.

We also analyse the frame-wise log-likehood ratios (based on MLE parameter estimates) comparing the Beta distribution to all other distribution, for both arousal and valence, and plot the histograms in Figures 3 and 4. Similarly, the histograms of frame-wise expected log-likelihood ratios comparing the Beta distributions to the others are shown in Figures 5 and 6. All four plots reveal that consistently, across majority of the data points or frames in the RECOLA dataset, the Beta distribution is a better choice than all the others.

6. Conclusion

Affective computing systems that model the distribution of numerical emotion ratings often assume a Gaussian distribution. In this paper we have shown that such an assumption is often incorrect. Seven of the most common distributions defined over a bounded domain were investigated as alternatives and the Beta distribution was found to be the most suitable to model numerical emotion labels. In particular, the Beta distribution consistently proved to be a better fit to the emotion ratings, across time, in the RECOLA dataset. Furthermore, while the RECOLA dataset was chosen for this study since it provides the highest number of label annotations amongst all the commonly employed datasets providing continuous affect annotations, the characteristics of continuous affect labels do not change with context. Therefore, the analyses and findings in this paper are expected to be relevant in all similar scenarios. As mathematical frameworks that consider ambiguity and uncertainty in affective computing systems develop, it can be expected that the choice of distribution to model affect labels would play an increasingly crucial role.

7. References

[1] R. Berrios, "What is complex/emotional about emotional complexity?" *Frontiers in psychology*, vol. 10, p. 1606, 2019.

[2] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment

research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[3] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2011.

[4] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[5] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–8.

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

[7] M. Schröder, L. Devillers, K. Karpouzis, J.-C. Martin, C. Pelachaud, C. Peter, H. Pirker, B. Schuller, J. Tao, and I. Wilson, "What should a generic emotion markup language be able to represent?" in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 440–451.

[8] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," 03 2017, pp. 5005–5009.

[9] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[10] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening," *Cognitive Computation*, pp. 1–10, 2020.

[11] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions using kalman filtering," in *2010 Ninth International Conference on Machine Learning and Applications*. IEEE, 2010, pp. 655–660.

[12] Y.-H. Yang and H. H. Chen, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2184–2196, 2011.

[13] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Modeling the affective content of music with a gaussian mixture model," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 56–68, 2015.

[14] T. Dang, V. Sethu, and E. Ambikairajah, "Dynamic multi-rater gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using kalman filters," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4929–4933.

[15] T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "An investigation of emotion prediction uncertainty using gaussian mixture regression." in *INTERSPEECH*, 2017, pp. 1248–1252.

[16] M. Atcheson, V. Sethu, and J. Epps, "Gaussian process regression for continuous emotion recognition with global temporal invariance," in *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*. PMLR, 2017, pp. 34–44.

[17] ——, "Demonstrating and modelling systematic time-varying annotator disagreement in continuous emotion annotation." in *Interspeech*, 2018, pp. 3668–3672.

[18] B. Zhang, E. M. Provost, R. Swedberg, and G. Essl, "Predicting emotion perception across domains: A study of singing and speaking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[19] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 3–10.

[20] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[21] A. Gelman *et al.*, "Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)," *Bayesian analysis*, vol. 1, no. 3, pp. 515–534, 2006.