Interspeech 2018
2-6 September 2018, Hyderabad

Demonstrating and modelling systematic time-varying annotator disagreement in continuous emotion annotation

Mia Atcheson, Vidhyasaharan Sethu, Julien Epps

University of New South Wales
mia.g.atcheson@gmail.com

Abstract

Continuous emotion recognition (CER) is the task of determining the emotional content of speech from audio or multimedia recordings. Training targets for machine learning must be generated by human annotation, generally as a time series of emotional parameter values. In typical contemporary CER systems and challenges, the mean over a pool of annotators is taken to represent this ground truth, but is this an appropriate model for the emotional content of speech? Using the RECOLA dataset, the primary contribution of this research is to show that a correlation exists between the time-varying disagreement from independent groups of annotators. Because the groups are completely isolated except via the speech signal, this agreement-about-disagreement demonstrates that there is a component of annotator disagreement which arises systematically from the signal itself, which qualitatively implies that the perceived emotional content of speech can exhibit some degree of inherent ambiguity. Additionally, we show that these human annotations exhibit a degree of temporal smoothness. Neither of these characteristics is represented by the standard series-of-means ground-truth model, so we propose two alternative ground-truth models: a mean-variance model that incorporates ambiguity, and a more general Gaussian process model that incorporates ambiguity and temporal smoothness in a well-defined probability distribution.

1. Background

1.1. Continuous emotion recognition

Continuous emotion recognition from speech is the task of taking an audio or multimedia recording of a human speaking, and, using the content of that recording, assigning numerical emotional parameter values to each temporal frame, in order to represent the emotional content of the speech signal as it varies through time [1]. In this way it is distinct from label-based schemes, which attempt to assign emotional labels from a finite set rather than real-valued parameters to describe affective content [2, 3], and from whole-utterance emotion recognition, which attempts to assign a single emotional description to the entire recording, rather than attempting to model its temporal development [4].

Numerous emotional parameter models have been proposed, with common parameters including *arousal* (level of excitement or predisposition to activity), *valence* (positive or negative attitude), and *dominance* (level of social dominance or submission communicated). These or other individual parameters can then be combined by placing them orthogonally in an emotional parameter space to create a dimensional model of the emotional content of speech [2, 5, 6, 7].

1.2. Annotation process

In order to apply machine learning techniques to the task of continuous emotion recognition, it is necessary to supply the predictive system with training data defined on that emotional parameter space. Because experimenters have no direct access to the emotions experienced by speakers and listeners, this training data is obtained through a process of human annotation [8, 9].

2. Characteristics of annotations

In order to model continuous emotion annotations in a principled way, it is relevant to analyze the various factors that determine the distribution of each annotator's response to the signal at each point in time. Consider a speech recording $\boldsymbol{S} = (\boldsymbol{S}[1], \boldsymbol{S}[2], \ldots, \boldsymbol{S}[N])$, with each frame $\boldsymbol{S}[t]$ a vector of values representing the content of the recording at that time (in practice, this may either be frames of the audio or multimedia file itself, or feature vectors derived therefrom), with corresponding annotations $\boldsymbol{a}[t] = (a_1[t], a_2[t], \ldots, a_M[t])$, where for each temporal frame index t^* and each of M annotators drawn from a population $\boldsymbol{P}$ (where $|\boldsymbol{P}|$ may be $>> M$), $a_m[t^*] \in [-1, 1]$ is the rating of annotator m at time t^*. There are a number of factors that could conceivably contribute to the value of $a_m[t^*]$:

1. The properties of the speech signal itself: $\boldsymbol{S}$
2. The annotation provided by that annotator at other times: $a_m[t]$ for $t \neq t^*$
3. Systematic factors specific to that particular annotator m.
4. Uncorrelated noise.

In the following sections, we will treat these human annotations as a random process, and analyze the effect on the distribution of $a_m[t^*]$ of the speech signal itself (1) and the annotations provided at other times (2). To isolate the effect of $\boldsymbol{S}$ on the distribution of $a_m[t^*]$, define

$$\hat{a}[t] = |\boldsymbol{P}|^{-1} \sum_{x \in \boldsymbol{P}} a_x[t]. \quad (1)$$

To describe the conditional distribution of $\hat{a}$ given $\boldsymbol{S}$ is the primary target of the prediction task of continuous emotion recognition: it is the distribution over the perceived emotional content of the utterance $\boldsymbol{S}$ at time t, averaged over the relevant population $\boldsymbol{P}$ from which the annotators $1 \ldots M$ are drawn. As mentioned above, the standard in the field of continuous emotion recognition is to use a series-of-means to represent the emotional ground truth, which corresponds probabilistically to a distribution with variance that is constant in time. As such, it is relevant to enquire: is the variance of the conditional distribution of $\hat{a}$ with respect to $\boldsymbol{S}$ indeed constant with respect to time, or does it have some meaningful time-varying nature which is not well-described by such a mean-only model?

10.21437/Interspeech.2018-1933

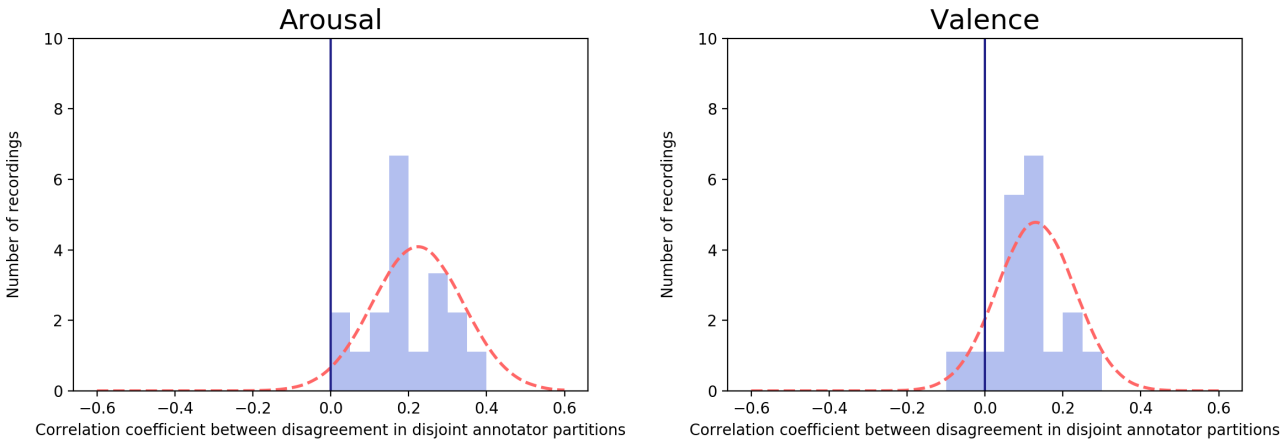

Figure 1: *Histogram of the Pearson correlation coefficient between the disagreement of annotator partitions on each recording in the dataset. The light blue bars represent the number of recordings with a correlation coefficient (averaged over all possible partitionings into two equal groups) in each of 20 histogram bins, and the dotted red line is a Gaussian distribution fit to this data. Note that aside from two valence recordings, 34 of the 36 total recordings exhibit a positive correlation.*

The variance of this conditional distribution corresponds to the inherent ambiguity of the speech signal itself, once unsystematic noise and listener-specific factors are excluded. Thus, we can rephrase the above question in the following terms: are some sections of speech inherently more ambiguous than others? Qualitatively, we can consider linguistic phenomena such as sarcasm: in English, an utterance such as *"Isn't that great!"* could be intended and perceived as conveying either praise or criticism, corresponding to either high or low valence, depending on whether the speaker is (perceived as) being genuine or sarcastic. As sarcasm is not always reliably perceived as intended, even within the same speaker-listener pair, this is an example of an utterance that could be considered to contain a high degree of inherent ambiguity. Emotional ambiguity has been investigated in a categorical emotion classification setting by Sobol-Shikler et al. [10], but has not recieved significant attention in a continuous, dimensional emotion recognition setting.

Regardless of the shape of the distribution at each individual point in time, there is also the matter of potential temporal interdependence in the joint distribution: is there some time-mediated relationship between the annotation values at different temporal frames? One important temporal property to consider is that of smoothness: is there some correlation between annotation values which are adjacent or nearby in time? Qualitatively, some degree of smoothness would seem appropriate, given the nature of emotion: if a person is angry now, it is likely that they will still be angry in 500ms. Annotators know this, and will not expect speech to change wildly in affect every frame, lending a natural smoothness property to their prior.

3. Experimental evaluation

3.1. Dataset

For the purposes of analyzing the properties of continuous emotion annotations, we use the RECOLA dataset [8]; specifically, the combined `train` and `dev` partitions of the data which are used in the AVEC 2017 affective computing challenge [11]. This dataset consists of 18 recordings of dyadic interactions in French, each lasting approximately 300 seconds, and annotated by six independent humans (with each recording using a new panel of annotators). This particular dataset was chosen as it is the only CER dataset of significant size and number of annotators available for which the researchers have access to the individual annotations, rather than only the mean over the annotators at each point in time, which was necessary to conduct analysis of annotator disagreement.

3.2. Ambiguity

We analyze the RECOLA emotion annotation dataset to attempt to distinguish whether such time-varying signal-dependent variability exists. First, the six human annotators $m = 1 \dots 6$ for each recording are partitioned into two groups of three (G_A and G_B), then the sample variances $\sigma_A[t]$ and $\sigma_B[t]$ over the three annotations are calculated for each temporal frame. Each of these two variance time series σ_A and σ_B are independent, and represent the level of disagreement amongst each group about the emotional content of the signal at each point in time. Some of this disagreement will arise from differences between the human annotators, and some will be due to unsystematic noise, but does any arise from the signal $\boldsymbol{S}$ itself? In order to analyze this, we can compare the shape of these two time series — if the signal is inherently unambiguous, and all disagreement arises from noise or from the differing characteristics of the human annotators, then we would not expect to see any correlation between the disagreement σ_A and σ_B of the two independent annotator groups G_A and G_B. However, if there is some systematic ambiguity in the speech signal, then we would expect to see some degree of correlation between the time series, as regions of high ambiguity in the speech signal should correspond to regions of high variance across both groups, and vice versa.

As detailed in Table 1, the Pearson correlation coefficient between the disagreement time series calculated from groups of three annotators, averaged over all possible annotator partitionings on the RECOLA dataset (AVEC2017 `train` and `dev` recordings) is 0.19 for arousal and 0.11 for valence. Intuitively, this is what we would expect to see if there was some degree of ambiguity inherent in the signal — a positive but small correla-

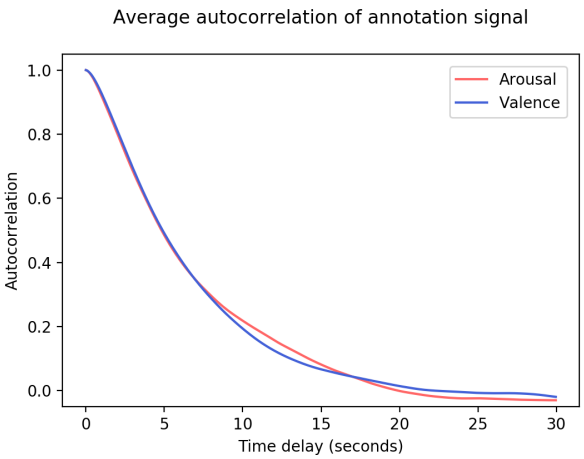

Figure 2: *Correlation coefficient between the annotation and a time-delayed version of itself, averaged over the 18 recordings.*

Table 1: *Results describing the correlation between the disagreement of independent annotator groups on the RECOLA dataset, showing the mean correlation coefficient across the 18 recordings, the sample standard deviation of the correlation coefficient, and the p-value for a one-tailed t-test, with null hypothesis that no correlation exists ($\mu = 0$).*

	Arousal	Valence
μ	0.19	0.11
σ	0.09	0.08
$p(\mu = 0)$	4.5×10^{-7}	4.9×10^{-5}

tion, with most of the annotator disagreement being attributable to noise or to annotator differences, but some portion arising from the signal itself. To demonstrate the statistical significance of these results, Table 1 also shows the p-values for a one-tailed t-test, with null hypothesis $\mu = 0$, which would be the case if no systematic ambiguity was present. For both arousal and valence, this p-value is less than 0.0001, indicating that these results would be extremely unlikely to be observed if no such correlation exists. For further illustration, Figure 1 shows a histogram of the correlation values for each of the 18 independent recordings in the dataset, along with a fitted Gaussian distribution.

3.3. Smoothness

We analyze the RECOLA emotion annotation dataset to determine the degree of correlation between annotated emotional parameter values as a function of their temporal distance. Figure 2 displays the Pearson correlation coefficient between the time series of mean RECOLA annotations (over the six annotators) with a time-delayed version of itself, averaged over the 18 recordings in the dataset. For both arousal and valence, this autocorrelation is high for the five seconds or so, but declines rapidly to zero at a 30-second delay, implying that the annotations display smoothness on the scale of a few seconds. This makes intuitive sense: we would not expect the emotional content of someone's speech to change significantly every second, but after half a minute, the emotional content may be very different.

4. Alternative models

In this section, we will discuss various models for the 'ground truth' continuous emotion recognition target for a single emotional parameter dimension. While the actual list of annotation values from each individual annotator contains all the available information, this may not be the optimal choice of representation for the ground truth. Such a model would grow linearly in complexity with the number of annotators, and so is probably not feasible for use with a large annotator pool. If we can choose an appropriate model that is able to more concisely summarize the relevant characteristics of the distribution of $\hat{a}$, the computational complexity of prediction can be reduced, and, if the representation of the information is more relevant to the prediction system, its accuracy could potentially be increased.

4.1. Mean-only

Currently, most continuous emotion recognition systems and challenges use a series-of-means representation of ground truth for each emotional parameter: at each point in time, the average annotation value is taken across the annotator pool, giving a mean-only ground-truth model $\mathcal{G}_{MO} = (\mu[1], \mu[2], \ldots, \mu[N])$ (see [11, 12, 13, 14]). To interpret this probabilistically, we can equip each temporal frame with a Gaussian distribution with mean μ_t and constant variance. While this model can represent the effect of the speech signal on the mean of the annotation distribution, it cannot represent a non-constant variance as discussed in Section 2, nor does it implement any temporal smoothness property as discussed in 2.

4.2. Mean-variance

The above mean-only model can be generalized into a mean-variance model by including a variance along with the mean at each temporal frame: $\mathcal{G}_{MV} = ((\mu[1], \sigma[1]), (\mu[2], \sigma[2]), \ldots, (\mu[N], \sigma[N]))$. This model is able to additionally represent non-constant variance as discussed in Section 2, and is thus able to model 'inherent ambiguity' in spoken utterances, unlike the previous model. Probabilistically, we can equip each temporal frame t with a Gaussian distribution $\mathcal{N}(\mu[t], \sigma[t])$. In practice, if these σ are derived from the sample variance of annotations, then it may be necessary to add a small constant value, or pass the series through some smoothing function, to avoid the degenerate case of zero variance.

4.3. Gaussian process

While the above mean-variance model is able to represent time-dependent variance in the distribution of annotations, it does not reflect any dependencies present between the annotation values at different points in time. In particular, if we were to draw a complete sample time series from a mean-variance distribution as described above, the result would be unlikely to display any temporal smoothness; a highly positive value (relative to the mean) is no less likely to occur after another highly positive value than after a highly negative value, which does not reflect an understanding of smooth short-term emotional development through time, which we have demonstrated experimentally in Section 3. In order to conform the distribution to our expectation of emotional content which varies at a limited rate in time, we can define the joint distribution over all points in time using a Gaussian process (GP). GPs provide a way to define a probability distribution over a function space by defining the covariance between output values as a function of the corresponding input

values [15]. If we consider the time index as an input value of the GP and the annotation (in emotional parameter space) as the output value of the function, then a condition of temporal smoothness can be applied to the distribution defined by the GP by using a covariance function which assigns higher covariance values to points which are closer in time. In a predictive setting, Atcheson et al. [16] showed that such a model could be combined additively with a covariance function defined over a feature space containing $\boldsymbol{S}$ to improve predictions over a non-temporal feature-space-only GP model. In order to use a GP in practice to represent a distribution over emotional annotations, the list of means in the previously-mentioned models can be indexed with a mean function $m[t] = \mu[t]$. The covariance function $c[t, t']$ can then be constructed as follows:

$$c[t, t'] = c_{\text{self}}[t, t'] + c_{\text{time}}[t - t'] \tag{2}$$

with $c_{\text{self}}[t, t'] = \sigma_t$ where $t = t'$, and 0 otherwise, and $c_{\text{time}}[t - t']$ represents relative temporal dependencies: the relationship between the distributions of annotations at differing points in time which is mediated by their temporal distance. To incorporate an assumption of smoothness, we can use a squared-exponential kernel:

$$c_{\text{time}}[t - t'] = \sigma^2 \exp\left(-\frac{(t - t')^2}{2l^2}\right) \tag{3}$$

where l determines the characteristic length-scale at which this smoothness assumption is applied. Equipped with mean function $m[t]$ and covariance function $c[t, t']$, samples from the Gaussian process $\mathcal{GP}(m[t], c[t', t])$ will be likely to display the required smoothness properties while still incorporating the information represented by the mean-variance model. This framework can also be extended to express more complex distributional features if necessary: mixture-of-GP models can incorporate multimodality in the shape of the distribution[17], while processes over non-symmetric distributions such as the skew-normal distribution can incorporate distributional asymmetry [18].

5. Discussion

While we have demonstrated that both ambiguity and temporal smoothness are present in the perceived emotional content of speech, this alone does not imply that it is necessary or desirable to adopt ground-truth models for CER that incorporate these features. However, there are a number of advantages of such models over the standard mean-only representation which may justify the additional complexity of these more general models:

- Qualitatively, these models may more correctly approximate the true emotional content of speech. Given that emotional communication and emotional expression are inherently ambiguous processes, an assumption of non-ambiguity as implied by mean-only ground-truth models runs counter to how we qualitatively understand paralinguistics.
- As this research has shown, time-varying ambiguity is a feature of the signal in its own right. Taking only the mean over the annotators destroys this information, so if a mean-only model is used as the training target for a machine learning system, the system is deprived of bona fide information about the training inputs, which it might otherwise be able to use to inform its predictions.
- The proposed models are strictly richer than the mean-only representation, so if used as the prediction target for a machine learning system, they may improve the utility of its output: if the predictions are used as input to another system, or are taken to be fused with the predictions of other systems to produce a multi-system fusion result, the predicted ambiguity may be useful to these downstream systems. For example, in a multi-system fusion, the ambiguity value can be treated as a measure of confidence in the prediction, and the contribution of the system to the multi-system fusion can be up-weighted or down-weighted based on the degree of ambiguity, even if the goal of the multi-system fusion is to predict only the mean value. Dang et al. in [19] demonstrate a system to predict annotator disagreement, and show that areas of lower disagreement correspond to more reliable predictions with a Gaussian mixture regression based system, suggesting that such an approach may improve predictions in a multi-system fusion setting.
- The proposed GP-based model is a well-defined probability distribution rather than simply a series of values, so it has a number of advantages: pointwise predictions can be directly compared against the distribution to determine their likelihood, natively probabilistic systems can consume the distribution directly, and samples can be taken from the distribution, which, if a temporal smoothness term is included in the GP covariance function, will display the relevant smoothness properties. Although the mean-only and mean-variance models can be equipped with a probabilistic interpretation which is useful on the scale of a single point in time, their lack of a temporal smoothness term prevents them from correctly modelling the properties of an entire annotation time series.

Ultimately, the utility of these alternative ground truth models would need to be established through practical usage. If future CER competitions were to include more detailed ground-truth models as targets alongside traditional mean-only models, this could provide a basis on which different approaches to ground-truth could be compared in practice.

6. Conclusion

In this paper, we have shown using the RECOLA dataset that there is a component of annotator disagreement in continuous emotional parameter annotation which arises from the speech signal itself, in addition to disagreement arising from differences between the human annotators or from unsystematic noise. We also show that these emotion annotations display a degree of temporal smoothness on the scale of a few seconds. Because standard series-of-means representations of ground truth emotional content used in typical contemporary CER applications are unable to model either of these features, we proposed two more general models: a mean-variance model that incorporates ambiguity, and a Gaussian process model that additionally incorporates temporal smoothness to produce a fully-fledged probability distribution over the perceived time-varying emotional content of an entire speech recording. These models offer a number of potential advantages over mean-only models: they more correctly describe our intuitive understanding of emotional communication, they are able to supply more information about training signals to machine learning systems, and predictions generated in this format may be more useful to downstream systems which consume those predictions.

7. Acknowledgements

This research was partly funded by an Australian government Research Training Program scholarship.

8. References

[1] H. Gunes, "Automatic, dimensional and continuous emotion recognition," 2010.

[2] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, 2008, pp. 597–600.

[3] L. F. Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cognition & Emotion*, vol. 12, no. 4, pp. 579–599, 1998.

[4] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[5] E. Schubert, "Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space," *Australian Journal of Psychology*, vol. 51, no. 3, pp. 154–165, 1999.

[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] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 827–834.

[8] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[9] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

[10] T. Sobol-Shikler and P. Robinson, "Classification of complex information: Inference of co-occurring affective states from their expressions in speech," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1284–1297, 2010.

[11] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.

[12] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, 2012.

[13] E. Pei, L. Yang, D. Jiang, and H. Sahli, "Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 208–214.

[14] 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*. ACM, 2016, pp. 3–10.

[15] C. K. Williams and C. E. Rasmussen, "Gaussian processes for machine learning," *the MIT Press*, vol. 2, no. 3, p. 4, 2006.

[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*, 2017, pp. 34–44.

[17] V. Tresp, "Mixtures of Gaussian processes," in *Advances in neural information processing systems*, 2001, pp. 654–660.

[18] M. Alodat and E. Al-Momani, "Skew Gaussian process for nonlinear regression," *Communications in Statistics-Theory and Methods*, vol. 43, no. 23, pp. 4936–4961, 2014.

[19] T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "An investigation of emotion prediction uncertainty using Gaussian mixture regression," *Proc. Interspeech 2017*, pp. 1248–1252, 2017.