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

Stochastic Process Regression for Cross-Cultural Speech Emotion Recognition

Mani Kumar T[1], *Enrique Sanchez*[1], *Georgios Tzimiropoulos*[2], *Timo Giesbrecht*[3], *Michel Valstar*[1]

[1]University of Nottingham, Nottingham, UK
[2]Queen Mary University London, London, UK
[3]Unilever R&D Port Sunlight, UK

{mani.tellamekala, michel.valstar}@nottingham.ac.uk, kike.sanc@gmail.com,
g.tzimiropoulos@qmul.ac.uk, timo.giesbrecht@unilever.com

Abstract

In this work, we pose continuous apparent emotion recognition from speech as a problem of learning distributions of functions, and do so using Stochastic Processes Regression. We presume that the relation between speech signals and their corresponding emotion labels is governed by some underlying stochastic process, in contrast to existing speech emotion recognition methods that are mostly based on deterministic regression models (static or recurrent). We treat each training sequence as an instance of the underlying stochastic process which we aim to discover using a neural latent variable model, which approximates the distribution of functions with a stochastic latent variable using an encoder-decoder composition: the encoder infers the distribution over the latent variable, which the decoder uses to predict the distribution of output emotion labels. To this end, we build on the previously proposed Neural Processes theory by using (a). noisy label predictions of a backbone instead of ground truth labels for latent variable inference and (b). recurrent encoder-decoder models to alleviate the effect of commonly encountered temporal misalignment between audio features and emotion labels due to annotator reaction lag. We validated our method on AVEC'19 cross-cultural emotion recognition dataset, achieving state-of-the-art results.

Index Terms: Cross-cultural Continuous Emotion Recognition, Stochastic Processes Regression, Neural Processes

1. Introduction

Recognising continuous-valued apparent emotions like valence (degree of pleasantness) and arousal (degree of activeness) [1], from speech signals is challenging [2], mainly because the processes of human emotion expression and perception are highly subjective [3]. Vocal emotion expressions are heavily influenced by factors like language tonality, sociocultural background, personality type, etc. [4]. Also, the perception of emotion is influenced by the cultural background of the observer [5, 6]. For this reason, while annotating emotion corpora it is a good practice to ensure that both the rater and the subject belong to the same culture [7].

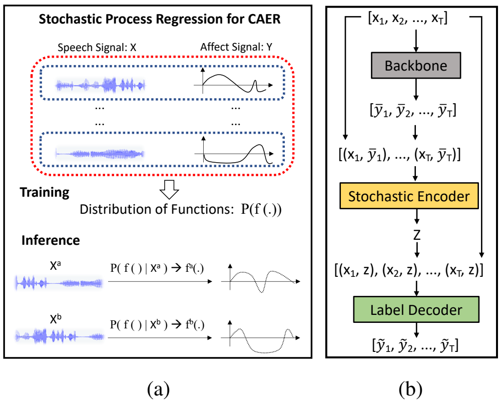
Stochastic Process Regression for CAER
Speech Signal: X
Affect Signal: Y
Training
Distribution of Functions: P(f (.))
Inference
X^a
P(f ( ) | X^a) → f^a(.)
X^b
P(f ( ) | X^b) → f^b(.)
[x1, x2, ..., xT]
Backbone
[ȳ1, ȳ2, ..., ȳT]
[(x1, ȳ1), ..., (xT, ȳT)]
Stochastic Encoder
z
[(x1, z), (x2, z), ..., (xT, z)]
Label Decoder
[ỹ1, ỹ2, ..., ỹT]
(a)
(b)

Figure 1: *(a) Learning a distribution over functions for continuous apparent emotion recognition and function sampling conditioned on the test inputs. (b) Stochastic Process Regression using Neural Processes (x_t — acoustic features, $\bar{y}_t$ — proxy labels, z — stochastic latent variable, $\tilde{y}_t$ — predicted labels).*

Most existing methods pose continuous apparent emotion recognition (CAER) from speech as a deterministic temporal regression problem, in which a single temporal function is learned. Methods based on RNNs (LSTMs and GRUs) [8, 9, 10] and 1D CNNs [11] applied to both low-level audio descriptors and deep audio representations, have shown promising performance on challenging CAER corpora in recent years. However, we argue that deterministic regression fundamentally lacks the flexibility to effectively model the intrinsic variations in the CAER labels. Prior work used domain adaptation [12, 13, 14, 15], transfer learning [16, 17], and meta-learning [18, 19, 20] methods to learn the variability of continuous emotion labels. Although these methods alleviate the problem to some extent, they still use deterministic regression and therefore we argue that they are inadequate to effectively capture the variations of emotion labels across different cultures.

We pose CAER from speech as a problem of learning a distribution over temporal functions i.e. Stochastic Process Regression ($\mathcal{SPR}$). We hypothesise that the following two key properties of $\mathcal{SPR}$ make it very appealing to CAER: (i) distributions over functions can capture the intrinsic variations of CAER more effectively than a deterministic function, inspired by the recent success of [21] and (ii) at test time $\mathcal{SPR}$ allows us to sample a function conditioned on the test input (Fig. 1a), thus making the model adaptable to the context of test inputs.

The family of Gaussian Processes (GPs) [22] is one of the widely used approaches to model the stochastic processes. GPs regression is applied to CAER problem in [23, 24], however, existing GPs regression methods applied to CAER are limited by two factors: the requirement for hand-designed kernel function and poor inference scalability (cubical complexity). To address these limitations a recent advancement in stochastic process modeling, the family of Neural Processes (NPs) [25, 26], effectively combines the representation learning capacity of deep neural networks (DNNs) with the inference time flexibility of Gaussian Processes (GPs). Unlike GPs, NPs implicitly learn a kernel function from the data and they are linearly scalable.

Similar to Variational Auto Encoders (VAEs) [27], NPs learn an implicit latent variable model using an encoder-decoder

http://dx.doi.org/10.21437/Interspeech.2021-610

model. In NPs, given a sequence of inputs and target labels, a subset of input-target pairs are used for stochastic context inference, to predict the distribution over labels of the remaining inputs. Direct application of original NPs implementation to CAER is constrained by (i). temporal misalignment between the input speech signals and target emotion labels sequence and (ii). NPs' requirement for the ground truth labels for the context frames to infer the latent variable at test time. Due to human annotator reaction lag [28], typically the emotion label sequence lags behind the speech signal by 2.5 to 4 seconds [11]. To alleviate the temporal misalignment effect, we use a temporal encoder-decoder model in the place of static encoder-decoder model. And to address the latter problem, we use proxy labels i.e. noisy predictions of a deterministic temporal model, in the place of ground-truth labels for the stochastic context inference.

As Fig 1b shows, our approach has three key components: i. Backbone — to generate noisy label predictions (proxy labels) from the input acoustic features, ii. Stochastic Encoder — to infer a stochastic latent variable from the acoustic features paired with the backbone predictions and iii. Label Decoder — to output predictive distributions over output label sequence. Here, the stochastic latent variable is supposed to capture a realisation of the underlying stochastic process conditioned on the test input sequence context. Leveraging the expressive power of stochastic processes, the proposed method achieves state-of-the-art results on AVEC19 Cross cultural Emotion recognition Sub-challenge (CES) corpus [9].

In summary our key contributions are:

- We pose continuous apparent emotion recognition from speech signals as a stochastic process regression problem to account for the intrinsic subjective nature of vocal emotion expression and perception.
- Our approach to stochastic process regression based on Neural Processes makes two key modifications to the original NPs implementation: proxy labels are used instead of ground truth labels for the stochastic latent variable inference and static encoder-decoder model is replaced with a recurrent encoder-decoder model.
- We validate our method on AVEC2019 CES dataset [9] and demonstrate state-of-the-art performance on cross-cultural apparent emotion recognition from speech data.

2. Methodology

Given a feature sequence of T audio frames, $X_t = [x_1, x_2, ..., x_T]$, our goal is to infer its affect label sequence, $Y_t = [y_1, y_2, ..., y_T]$. Due to this task's intrinsic temporal nature, it is generally assumed that the target label y_t is a function of not only its corresponding feature x_t but also the temporal context h_t that is distributed over the frames $[x_{t-N}, .., x_{t+N}]$.

One natural model choice to learn this temporal task is a recurrent neural network (RNN), $Y_t = f_r(X_t, h_t)$. Most existing CAER methods typically use LSTM-RNNs or GRU-RNNs [29, 11, 30] to learn a single deterministic temporal function from the training sequences. We argue that a single deterministic function has limited representation capacity to effectively model the intrinsic variations (e.g. culture- and person-specific differences) in the continuous apparent emotions.

Continuous Apparent Emotion Recognition as a Stochastic Process. We propose to model the latent variables of CAER (e.g. culture-specific or person-specific factors) as random variables whose probability distributions are drawn from a *Stochastic Process* ($\mathcal{SP}$). We presume that for an input feature sequence X_t, the target emotion label sequence Y_t is a function of not only X_t but also a stochastic latent variable Z which is approximated by a random variable with a multivariate normal distribution $\mathcal{N}(\mu_Z^i, \sigma_Z^i)$, drawn from the underlying $\mathcal{SP}$.

$$Y_t = f_{\mathcal{SP}}(X_t, Z), \tag{1}$$

2.1. Stochastic Process Regression using Neural Processes

Given the input acoustic feature sequence X_t, we first encode the stochastic context Z. Then the inferred context Z in combination with the feature sequence X_t is used to predict the distribution over output labels $\tilde{Y}_t$, following Neural Processes [26].

Stochastic Context Encoding (f_{enc}). As Fig. 1b shows, the stochastic latent variable Z is a function of two variables: input features and task labels. Here we use proxy labels $\bar{Y}_t$ as the task labels, where $\bar{Y}_t$ are predictions of a deterministic temporal model which we refer to as *backbone*. The proxy label sequence $\bar{Y}_t$ along with the input feature sequence X_t is fed into the encoder to estimate the latent variable distribution parameters (μ_Z^i, σ_Z^i). Note that the original implementation of NPs [26] requires ground truth labels for the context inference, which seriously limits the use cases in which one can apply NPs to CAER.

Decoding Output Label Distribution (f_{dec}). To predict the distributions over output emotion labels, similar to NPs, we first sample a vector Z from the latent variable distribution $\mathcal{N}(\mu_Z^i, \sigma_Z^i)$. As shown in Fig. 2, the decoder takes as input the sampled latent vector Z that is appended frame-wise to the input feature sequence X_t, and it outputs the parameters of distributions $(\mu_{Y_t}, \sigma_{Y_t})$ over the predicted label sequence.

In summary, the functional composition of $\mathcal{SPR}$-NPs is,

$$\begin{aligned}\tilde{Y}_t = f_{\mathcal{SPR}-NP}(X_t, Z) &= f_{dec}(X_t, f_{enc}(X_t, \bar{Y}_t)) \\ &= f_{dec}(X_t, f_{enc}(X_t, f_{\mathcal{BB}}(X_t))),\end{aligned} \tag{2}$$

where f_{enc} is the global context encoder, f_{dec} is the output label distribution decoder, and $f_{\mathcal{BB}}$ is the backbone model — a deterministic temporal function to generate the proxy labels $\bar{Y}_t$.

Realisation of $\mathcal{SPR}$-NP. The static encoder and decoder functions of the original NP implementation are not suitable to CAER given that the audio feature and label sequences are not temporally aligned due to annotator reaction lag [28]. Hence we use a RNN encoder-decoder model instead of a multi-layer perceptron encoder-decoder model used by NPs. As Fig. 2 illustrates, we transform the acoustic features and their proxy labels into embedding vectors using feature-label embedding networks. Then a RNN-encoder takes as input the concatenated feature and label embedding vector sequence to derive a context summary vector which is further mapped to the stochastic latent variable distribution parameters (μ_Z, σ_Z). A RNN-decoder then takes as input the sampled latent vector Z paired with the input feature sequence X_t to infer the predicted label sequence distribution parameters $(\mu_{Y_t}, \sigma_{Y_t})$. Note that we use the mean values μ_{Y_t} of the output label distribution as the final predictions for our performance evaluation.

Network Architectures. Fig. 2 depicts input-output dimensions of different modules of our $\mathcal{SPR}$-NP network. This network is primarily composed of a backbone net — a 4-layer bidirectional GRU (BiGRU-RNN) model with 128 hidden units, feature and label embedding nets — two 3-layer fully connected networks with 64 hidden units, a RNN encoder net — a 2-layer BiGRU-RNN with 64 hidden units and a RNN decoder net — a 2-layer BiGRU-RNN with 128 hidden units.

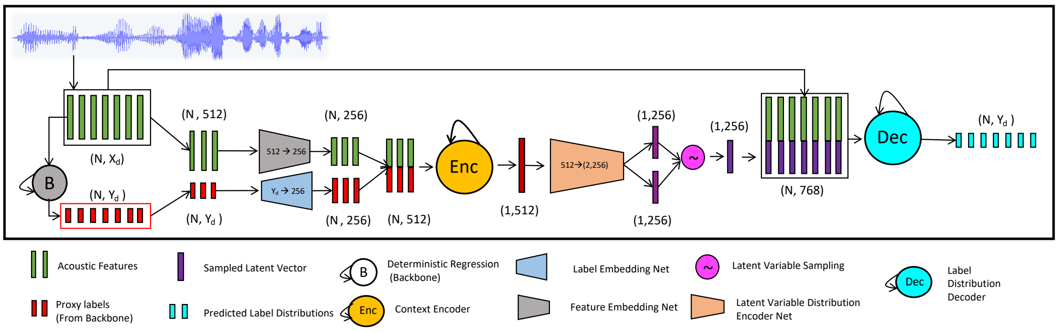

Figure 2: *$\mathcal{SPR}$-NP network architecture (N:number of input frames, X_d and Y_d:number of feature and label dimensions respectively)*

Training. Similar to NPs, we train the $\mathcal{SPR}$-NPs to minimise the Evidence Lower Bound (ELBO) objective function [27]:

$$L_{ELBO} = \mathbb{E}_{q(z)}(-\log p(Y_t|X_t, Z)) + KL(q_{cont}||q_{targ}) \tag{3}$$

using the reparameterisation trick [27], where $p(Y_t|X_t, Z)$ is the likelihood of ground truth labels Y_t according to the decoder output predictive distribution $\mathcal{N}(\mu_{Y_t},\sigma_{Y_t})$, q_{cont} is the encoder output latent distribution $\mathcal{N}(\mu_z^c,\sigma_z^c)$ for the context frames, and q_{targ} is the encoder output latent distribution $\mathcal{N}(\mu_z^t,\sigma_z^t)$ for the target frames. Note that during training we use all frames of the input sequence X_t as the target frames. But as context frames only a sub-set of the input frames are used in order to induce variability in the learned stochastic context. We vary the length of the context frames sequence (composed of consecutive frames from the input sequence) to be between 10 and the input sequence length using uniform random sampling. Note that the backbone model is trained first and its weights are frozen during $\mathcal{SPR}$-NP training, as the joint training of backbone and $\mathcal{SPR}$-NP is found to be unstable.

Inference. During testing, we use all frames in the input sequence as the context frames to infer the latent variable distribution. Although the decoder estimates both the mean and variance values for the output labels, here we consider only the mean values for model evaluation, leaving the variance evaluation and its role in uncertainty modelling for future work.

3. Experiments

We evaluated the proposed method on the AVEC 2019 Cross-culture Emotion Sub-Challenge (CES) [9] dataset. We trained the $\mathcal{SPR}$-NP models and compared their performance with different deterministic regression (DR) baselines.

Dataset. The AVEC 2019 CES dataset [9] contains a subset of SEWA [7] of audio-visual recordings of interactions between pairs of individuals from German, Hungarian and Chinese cultures. As the labels of test sets and Chinese culture are not publicly available, we used the German and Hungarian training and validation sets. Audio data is recorded at 48 kHZ and the ratings of valence and arousal are presented at 10 FPS. Liking dimension of this dataset is not used in this work as the liking recognition typically needs linguistic features that are explicitly derived [8], whereas we focus only on the audio-modality here. To demonstrate the generality of our method, we evaluated $\mathcal{SPR}$-NPs using both hand-crafted and deep audio features. Following Chen et al. [30], we used sequences of 30 seconds for training (with zero padding applied to shorter sequences).

Hand-crafted Acoustic Features. Using the OpenSMILE toolkit [31], we extracted two sets of hand-designed acoustic features: eGeMAPS [32] and Mel Frequency Cepstral Coefficients (MFCCs). Following prior work [8, 9, 10, 11], we computed first and second order functionals of both feature sets using a sliding window of 4 seconds with a stride of 100 ms. This results in 46 dimensional feature vectors for eGeMAPS and 78 dimensional feature vectors for MFCCs produced at 10 FPS.

The feature vectors are normalised with respect to the training set statistics separately for each data partition. Since in each audio file both speakers are present, a binary feature is prepared appended to the feature vectors, to indicate the presence of target speaker by exploiting the turn-taking information [11].

Deep Acoustic Features. In the AVEC19 CES, deep representation learning methods [30] demonstrated significantly better performance than the hand-crafted features [9]. Hence, we adopted the audio feature learning method proposed in [30] for evaluating the $\mathcal{SPR}$-NP model. We extracted 128-dimensional features by applying VGGish [33] pre-trained network to Mel-spectrogram images of the input audio signals (hop size and window length values set to 0.1s and 1s respectively). To differentiate the target speaker's features from the interlocutor's features, we followed dimensionality-doubling strategy proposed in [34]. Note that during training we fine-tuned only the last two fully connected layers of the VGGish pre-trained network.

Table 1: *Results (valence CCC, arousal CCC) using VGGish features and comparison with state of the art methods on AVEC19 CES dataset. (†denotes in-house implementation)*

Method	German	Hungarian	Ger.+Hun.
RUC [15]	(0.542,0.577)	(0.098,0.385)	(0.388,0.518)
NPU [30]	(0.554,0.622)	(0.143,0.408)	(0.435,0.562)
BiGRU†	(0.515,0.636)	(0.196,0.367)	(0.421,0.568)
CNP [25]†	(0.534,0.642)	(0.182,0.439)	(0.419,0.584)
$\mathcal{SPR}$-NP	(**0.576**,**0.671**)	(**0.209**,**0.492**)	(**0.441**,**0.618**)

Evaluation Metric. We used Lin's Concordance Correlation Coefficient (CCC) [35] to measure the CAER performance, following the AVEC CES protocol [8, 9]. CCC is a robust statistical measure that includes both precision and accuracy in a single metric that is not biased by scale and location variations.

Deterministic Regression (DR) Baselines. We trained a 4-layer bidirectional GRU-RNN (BiGRU-RNN) with 128 hidden units and with a dropout value of 0.5, as a representative model of the deterministic regression methods. To train the BiGRU-RNNs, we used a hybrid loss function, $L = L_{MSE} + L_{iCCC}$ where $L_{iCCC} = 1 - CCC(Y_{gt}, Y_{pred})$ and L_{MSE} is the mean square error between Y_{gt} and Y_{pred}. Note that this BiGRU-

Table 2: *Results (valence CCC, arousal CCC) using hand-crafted features on AVEC19 CES; BiGRU (bidirectional GRU) denotes the deterministic regression, and $\mathcal{SPR}$-NP denotes the stochastic process regression using NPs (†the audio baseline results reported in [9])*

		On eGeMAPS Functionals			On MFCCs Functionals		
		Evaluation Culture			Evaluation Culture		
Train Culture	**Model**	**German**	**Hungarian**	**Germ.+Hung.**	**German**	**Hungarian**	**Germ.+Hung.**
German	BiGRU	(0.422,0.519)	(0.032,0.221)	(0.275,0.419)	(0.349,0.446)	(0.031,0.295)	(0.193,0.378)
German	$\mathcal{SPR}$-NP	(**0.459**,**0.544**)	(**0.111**,**0.296**)	(**0.315**,**0.460**)	(**0.420**,**0.517**)	(**0.042**,**0.349**)	(**0.210**,**0.408**)
Hungarian	BiGRU	(0.019,0.279)	(0.103,**0.265**)	(0.052,0.289)	(0.141,0.214)	(**0.049**,0.263)	(0.104,0.239)
Hungarian	$\mathcal{SPR}$-NP	(**0.260**,**0.349**)	(**0.159**,0.262)	(**0.183**,**0.334**)	(**0.267**, **0.398**)	(0.023,**0.268**)	(**0.135**,**0.321**)
Germ.+Hung.	AVEC19† [9]	(0.405,0.396)	(0.073,0.305)	(0.286,0.371)	(0.344,0.389)	(0.017,0.236)	(0.187,0.326)
Germ.+Hung.	BiGRU	(0.377,0.449)	(0.136,0.308)	(0.306,0.410)	(0.351,0.419)	(0.044,0.317)	(0.212, 0.374)
Germ.+Hung.	$\mathcal{SPR}$-NP	(**0.420**,**0.471**)	(**0.144**,**0.316**)	(**0.327**,**0.428**)	(**0.414**,**0.525**)	(**0.058**,**0.344**)	(**0.235**,**0.436**)

Table 3: *Null-vector ablations using VGGish features on Germ.+Hung. evaluation set. Full $\mathcal{SPR}$: Complete $\mathcal{SPR}$-NP model, null-Z: with zero latent vector passed to the decoder, null-$\bar{Y}_c$: with zero proxy labels passed to the encoder and null-X_c: with zero feature vectors passed to the encoder.*

Full $\mathcal{SPR}$	With null-Z	With null-$\bar{Y}_c$	With null-X_c
(0.441,0.618)	(0.149,0.18)	(0.309,0.484)	(0.38,0.534)

RNN model is also used as the backbone model to source the proxy labels for the training and evaluation of $\mathcal{SPR}$-NP.

Along with the BiGRU-RNNs, we evaluated an additional DR baseline: Conditional NPs [25], a deterministic variant of NPs without the stochastic latent variable. Training and evaluation procedures that we followed for CNPs are same as that of $\mathcal{SPR}$-NPs except that the CNP encoder network outputs only a deterministic context vector and negative log-likelihood is used in the place of ELBO as the training objective for CNPs.

Optimisation Details. To learn the network weights of the DR baselines and $\mathcal{SPR}$-NP models, we used an Adam optimiser [36] with the values of batch size, initial learning rate and weight decay set to 16, 0.0001 and 0.00001 respectively. Learning rate is tuned using cosine annealing with warm restarts [37]. All the models are trained for 500 epochs using early stopping.

3.1. Results

As Table 1 shows, $\mathcal{SPR}$-NP model outperformed the deterministic regression baselines, both BiGRU-RNNs and CNPs on all three evaluation sets: German, Hungarian, and German+Hungarian. Note that these models are trained on German+Hungarian training set using VGGish features. Though both $\mathcal{SPR}$-NP and CNPs have almost the same number of weights, yet $\mathcal{SPR}$-NP achieved better results due to the flexibility rendered by its stochastic latent variable. $\mathcal{SPR}$-NP also achieved state-of-the-art results, outperforming the AVEC19 CES challenge winner [15], that uses adversarial domain adaptation, and runner-up [30] methods (in the audio modality category). These results demonstrate that the additional representation capacity of stochastic processes can enhance the generalisation performance of CAER models across different cultures.

Table 2 presents the results of BiGRU-RNN baselines and $\mathcal{SPR}$-NP models trained using the hand-crafted audio features. Here we trained culture-specific models, as the hand-crafted features typically require fewer training sequences than the deep audio features based methods. Similar to the results of VGGish features in Table 1, $\mathcal{SPR}$-NP models improved over the BiGRU-RNNs and AVEC19 CES baselines in almost all cases. Particularly, the performance gains achieved by $\mathcal{SPR}$-NP are significant in the cross-cultural evaluation protocol i.e. trained on German and evaluated on Hungarian or vice versa.

To understand the significance of stochastic context encoder variables, we performed null-vector ablations: during inference we pass a null-vector as input instead of the original vector of a variable to be evaluated. Note that the performance drop with null-vector inputs used only at test time, is not obvious in the latent variable models given their vulnerability to posterior collapse [38] which degrades the latent vector quality. We conducted these ablations on each of the following encoder variables separately: stochastic latent vector Z, context feature vectors X_c, and proxy label vectors $\bar{Y}_c$. The results (Table 3) confirm that the stochastic latent vector significantly contributes to the overall performance gains achieved by $\mathcal{SPR}$-NPs.

4. Conclusion

To account for variations like culture- and person-specific differences in vocal emotion expression and perception, we posed continuous apparent emotion recognition from speech as a problem of learning distributions over functions. Building on Neural Processes, we proposed a method for CAER from speech by addressing two major limitations of the original NPs implementation. Our method achieved state-of-the-art results on a challenging cross-cultural apparent emotion recognition database, capitalising on the expressive power of stochastic processes. In addition to the point estimates of apparent emotion labels, our model also outputs uncertainty estimates. We leave the assessment of uncertainty estimates quality and their applications in downstream affect computing tasks for future work.

5. Acknowledgements

The work of Mani Kumar T was supported by the Engineering and Physical Science Research Council project (2159382) and Unilever U.K. Ltd. The work of Michel Valstar was supported by the Nottingham Biomedical Research Centre (BRC).

6. References

[1] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, p. 715, 2005.

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[3] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *arXiv preprint arXiv:1909.00360*, 2019.

[4] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis." *Psychological bulletin*, vol. 128, no. 2, p. 203, 2002.

[5] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proceedings of the National Academy of Sciences*, vol. 107, no. 6, pp. 2408–2412, 2010.

[6] 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.

[7] 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 TPAMI*, 2019.

[8] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *AVEC*, 2018, pp. 3–13.

[9] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *AVEC*, 2019, pp. 3–12.

[10] A. Mallol-Ragolta, N. Cummins, and B. W. Schuller, "An investigation of cross-cultural semi-supervised learning for continuous affect recognition," *Proc. Interspeech 2020*, pp. 511–515, 2020.

[11] M. Schmitt, N. Cummins, and B. W. Schuller, "Continuous emotion recognition in speech—do we need recurrence?" *Proc. Interspeech 2019*, pp. 2808–2812, 2019.

[12] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *ICASSP*. IEEE, 2015, pp. 5058–5062.

[13] ——, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *ICASSP*. IEEE, 2017, pp. 5000–5004.

[14] ——, "Incremental adaptation using active learning for acoustic emotion recognition," in *ICASSP*. IEEE, 2017, pp. 5160–5164.

[15] J. Zhao, R. Li, J. Liang, S. Chen, and Q. Jin, "Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions," in *AVEC*, 2019, pp. 37–45.

[16] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, 2017.

[17] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE TAC*, vol. 10, no. 1, pp. 85–99, 2017.

[18] R. Cai, K. Guo, B. Xu, X. Yang, and Z. Zhang, "Meta multi-task learning for speech emotion recognition," *Proc. Interspeech 2020*, pp. 3336–3340, 2020.

[19] T. Fujioka, T. Homma, and K. Nagamatsu, "Meta-learning for speech emotion recognition considering ambiguity of emotion labels," *Proc. Interspeech 2020*, pp. 2332–2336, 2020.

[20] A. Naman and L. Mancini, "Fixed-maml for few shot classification in multilingual speech emotion recognition," *arXiv preprint arXiv:2101.01356*, 2021.

[21] E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos, "Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition," in *CVPR*, 2021.

[22] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer school on machine learning*. Springer, 2003, pp. 63–71.

[23] 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.

[24] ——, "Using gaussian processes with lstm neural networks to predict continuous-time, dimensional emotion in ambiguous speech," in *ACII*. IEEE, 2019, pp. 718–724.

[25] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. A. Eslami, "Conditional neural processes," in *ICML*, 2018, pp. 1704–1713.

[26] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. M. A. Eslami, and Y. W. Teh, "Neural processes," in *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[28] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *ACII*. IEEE, 2013, pp. 85–90.

[29] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*. IEEE, 2016, pp. 5200–5204.

[30] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli, "Efficient spatial temporal convolutional features for audiovisual continuous affect recognition," in *AVEC*, 2019, pp. 19–26.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[32] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE TAC*, vol. 7, no. 2, pp. 190–202, 2015.

[33] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *ICASSP*. IEEE, 2017, pp. 131–135.

[34] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *AVEC*, 2017, pp. 19–26.

[35] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[38] J. Lucas, G. Tucker, R. Grosse, and M. Norouzi, "Understanding posterior collapse in generative latent variable models," 2019.