Interspeech 2024
1-5 September 2024, Kos, Greece

Can Modelling Inter-Rater Ambiguity Lead To Noise-Robust Continuous Emotion Predictions?

Ya-Tse Wu[1], *Jingyao Wu*[2], *Vidhyasaharan Sethu*[2], *Chi-Chun Lee*[1]

[1]National Tsing Hua University, Taiwan
[2]University of New South Wales, Australia

crowpeter@gapp.nthu.edu.tw, jingyao.wu@unsw.edu.au, v.sethu@unsw.edu.au, cclee@ee.nthu.edu.tw

Abstract

There has been increasing attention drawn to modelling inter-rater ambiguity in Continuous Emotion Recognition (CER) systems using probability distributions for arousal and valence. However, the relationship between modelling label ambiguity and robustness to noise, and more broadly, the impact of real-world noise on CER systems remains insufficiently explored. In this study, we argue that incorporating inter-rater ambiguity during training can regularize the noise response, leading to noise robustness. To this end, we propose a novel loss function that incorporates inter-rater ambiguity into model training. Experiments conducted on the RECOLA dataset demonstrate that our proposed method achieves a maximum Concordance Correlation Coefficient (CCC) improvement of 0.117 and 0.077 for mean and standard deviation predictions, respectively, across all noise conditions. We further integrate traditional noisy augmentation strategies with our proposed method and observe promising results.

Index Terms: Continuous emotion prediction, Inter-rater ambiguity, Noise robustness

1. Introduction

Speech emotion recognition (SER) plays a pivotal role in developing natural human-computer interaction. Within the SER community, the complexity and richness of real-world emotions has led them to be represented with time and value-continuous affect dimensions, such as arousal and valence [1], leading to the development of Continuous Emotion Recognition (CER). The continuous emotion labels are typically collected from multiple raters who listen to audio recordings and provide their annotations within a certain numerical range, resulting in time-series ratings. However, inherent differences in perception, known as 'inter-rater ambiguity', exist due to differences in the perception of different individuals. This ambiguity inherently captures the richness of emotional nuances present in real-world speech and should be considered in the system development.

However, conventional CER systems often treat ambiguity as unwanted 'uncertainty' and only model the mean of the ratings [2, 3]. Recently, there has been a growing recognition of the importance of modeling ambiguity and developing ambiguity-aware emotion recognition systems, wherein ambiguity is modelled with probability distributions. For instance, Wu et al. developed a Sequential Monte Carlo framework as a non-paramtric and non linear dynamical model for predicting ambiguous emotion states [4]. Atcheson et al. utilised Gaussian processes to capture the label ambiguity with Gaussian distributions and employed a long short-term memory (LSTM) networks to make predictions [5]. Additionally, Bose et al. demonstrated the effectiveness of employing a Beta distribution for ambiguity modelling and utilized a similar neural network structure for predicting ambiguous emotion states [6]. These studies have expanded the feasibility of modelling the ambiguity and subtlety of emotions, setting the stage for modelling emotion in more complex scenarios of real-world applications, such as those encountered under noisy conditions.

In the real-world applications, the background noises degrade the quality of the original speech signals, leading to the degradation of SER performance. Despite extensive efforts to enhance the noise robustness of SER for traditional emotion classification tasks through various techniques such as speech enhancement [7, 8] or noisy data augmentation [9, 10], the impact of noise on continuous emotion prediction systems remains inadequately explored, particularly for CER systems that are ambiguity-aware. Given that humans exhibit differences in the perception of emotion, could this diversity lead to a robustness to noise? In particular, it raises the question: Can modelling emotion perceptions from multiple raters (i.e., inter-rater ambiguity) enhance the noise robustness of CER systems?

An SER system's learning process involves rater subjectivity and is vulnerable to noise-induced signal variability, both of which can impact its performance. Under these dual-factors variability idea, in this paper, we hypotheses that the CER system trained with proper rater subjectivity help constraint the model behavior to be robust even when faced with noise (i.e., signal variability). Specifically, we investigate the noise robustness of CER systems when inter-rater ambiguity is incorporated and modelled with probability distributions. Furthermore, we propose a novel technique to integrate ambiguity into system training through a loss function. Leveraging the advantages of the recently proposed metric, the 'Belief Mismatch Coefficient (BMC)', which offers both quantitative and interpretable comparison by directly assessing predicted distributions against ground truth ratings [11], we propose a novel 'Belief Mismatch Loss (BML)'. It measures the loss in belief that emotion states belong to a given region under both predicted and inferred distribution derived from the multi-rater labels. This loss effectively captures the ambiguity into system learning as opposed to conventional loss functions such as concordance correlation coefficient and mean squared error, aiming to further enhance the CER system's capability in noisy conditions.

2. Methodology

2.1. Target Distribution Parameterization

Given a set of ground truth ratings $\boldsymbol{y_n} = \{y_{n,1}, y_{n,2}, ..., y_{n,m}\}$ from m raters at a single time instance n, we assume the emotion states with associated ambiguity could be modelled as a Beta distribution since it has been shown to be the most suitable candidate to capture inter-rater ambiguity in CER tasks [12]. Following a similar approach to previous studies [13], the target Beta distribution $\boldsymbol{\beta_n}$ are derived from ground truth ratings

10.21437/Interspeech.2024-482

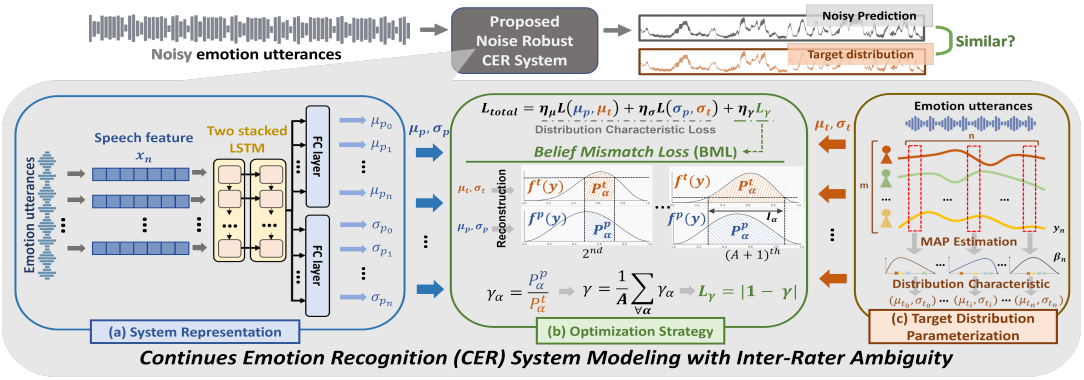

Figure 1: *Illustration of our framework and hypothesis.*

$\boldsymbol{y_n}$ as a Maximum A Posterior (MAP) estimate. The most commonly utilised parameterization pair of Beta distribution are the either the shape parameters $\{a, b\}$, {mean (μ), standard deviation (σ)} or {mode, concentration} [13]. For a better interpretation of the distribution characteristics, we adopt $\{\mu, \sigma\}$ pair as our Beta parameters, and a direct conversion between $\{a, b\}$ and $\{\mu, \sigma\}$ are given as: $\mu = \frac{a}{a+b}, \sigma = \frac{ab}{(a+b)^2(a+b+1)}$. Here, a, b must satisfy the constraint that $a > 1$ and $b > 1$ for a bell-shaped Beta distribution to be valid for capturing ambiguity in continuous emotion states.

2.2. System Representation

2.2.1. System Model Structure

A graphical representation of the system architecture is illustrated in Figure 1. A multi-task learning strategy is employed for predicting the time-varying Beta distributions $\boldsymbol{\beta} = \{\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_n}\}$ from the input speech features $\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}\}$ under different noisy levels. Following the backbone model structure of the SOTA framework [13], two stacked Long Short-Term Memory (LSTM) layers shared their weights in the front-end of the model and two branches of fully connected layers with a sigmoid activation function are implemented. The Beta parameter mean and standard deviation (SD) are simultaneously predicted in parallel.

2.2.2. Optimization Strategy

During system training, the goal is to minimize discrepancies between predicted and target distributions. To incorporate inter-rater ambiguity into the learning process, we introduce a novel loss function that assesses the predicted distribution against the target distribution from two perspectives: 1) measuring deviations in typical distribution characteristics like mean and standard deviation (SD); 2) evaluating overall distribution shape mismatch. The proposed loss function is formulated as follows.

$$L_{total} = \eta_\mu L(\boldsymbol{\mu}_p, \boldsymbol{\mu}_t) + \eta_\sigma L(\boldsymbol{\sigma}_p, \boldsymbol{\sigma}_t) + \eta_\gamma L_\gamma \quad (1)$$

where $\boldsymbol{\mu}_*$ and $\boldsymbol{\sigma}_*$ denote the mean and standard deviation (SD) sequences of either predicted (p) or target (t) distributions . η_* is the adjustable loss constant weight for each loss term. $L(.)$ refers to the concordance correlation coefficient (CCC) loss, which commonly utilized for time-series predictions and also adopted in ambiguity-aware CER systems [14, 13].

We propose the belief mismatch loss (BML), denoted as L_γ in Equation 1, as a means to incorporate a nuanced awareness of ambiguity into model training. Inspired by the belief mismatch coefficient (BMC) [11], a recent metric for measuring the accuracy of ambiguous emotion predictions, BML quantifies the mismatch between the distributions by comparing the belief that emotion states belong to a given region. Details of BML are further elaborated in the subsequent section.

2.3. Proposed Belief Mismatch Loss

2.3.1. Belief Mismatch Coefficient (BMC)

Belief Mismatch Coefficient (BMC) is a novel metric, gauging the prediction accuracy of an ambiguity-aware emotion prediction system. A value of 1 signifies a perfect match between the probability density function (PDF) of predicted ($f^p(y)$) and inferred distributions ($f^t(y)$) , quantifying the mismatch in belief about whether emotion states fall within a specific range [11].

In the original BMC approach, an α-likely region (I_α) is defined as the area formed by the two intersection points of any horizontal line and the target PDF (Refer to 1(b)), which is used to compute the belief mismatch ratio (BMR, γ_α) via per Equation 2. The BMR indicates how well the arousal/valence region predicted by the system, with probability P_α, matches the region inferred from the ground truth ratings.

$$\gamma_\alpha = \frac{P^p_\alpha}{P^t_\alpha}, P^*_\alpha = \int_{I_\alpha} f^*(y)dy \quad (2)$$

where P^*_α indicates the belief that the perceived emotion states fall within I_α under predicted or inferred distribution $f^*(y)$.

The BMR across all possible α-likely regions are then computed, from the most likely to the least likely regions. Finally, the BMC is computed as:

$$\gamma = \frac{1}{A}\sum_{\forall\alpha}\gamma_\alpha \quad (3)$$

where A denotes the number of different intervals (I_α) for which BMRs were estimated.

2.3.2. Belief Mismatch Loss (BML)

Inspired by BMC, we propose the *Belief Mismatch Loss* (BML), denoted as L_γ. Figure 1 (b) illustrates the idea of BML. Given a predicted distribution $f^p(y)$ (blue) and the target distribution $f^t(y)$ (orange) inferred from the ground truth ratings, the BML measures the loss that the belief of the emotion states falling into a region along the affect dimension (arousal/valence) under $f^p(y)$ against that under $f^t(y)$.

We define $A + 2$ regions of I_α by dividing the interval $[0, \max(f(y))]$ into $A + 2$ equal steps. The BMR is then computed according to Equation (2) at each region from the 2^{nd} to the $(A + 1)^{th}$ region. This process is depicted in Figure 1 (b), where the horizontal line loops from the top to the bottom. The first and last regions are excluded because when I_α approaches zero or encompasses the entire arousal/valence y plane, the beliefs on both the predicted and target distributions are either close to 0 or equal to 1, making them irrelevant for the purpose of comparisons. Finally, the BML is computed from BMC (γ) as:

$$L_\gamma = |1 - \gamma| \quad (4)$$

Although the original BMC approach is not constrained to only Beta distributions, it is only valid for unimodal distributions. However, as it is now adapted into a loss function within the system training process, the predicted distributions may take on arbitrary shapes. Therefore, the BML is computed only for those predictions that satisfy the criteria of a unimodal Beta distribution, as constrained by $a > 1$ and $b > 1$. That is, $\eta_\gamma = 0$ for non-bell-shape data points; otherwise, $\eta_\gamma = 1$ in Equation (1). We provide the implemented BML code on github [1].

3. Experimental Setup

3.1. Speech Emotion Dataset

RECOLA [15] is a widely used French multimodal corpus for continuous emotion recognition tasks, especially ambiguity modeling [4]. It consists of 9.5 hours of spontaneous conversation with each utterance spanning 5 minutes. Following the split rule in the Audio/Visual Emotion Challenge and Workshop (AVEC) 2016 [16], the utterances number are 9:9 in training and development sets, respectively. Since the original testing set is not publicly available, we recompose the a ratio of 8:2:8 for training, validation, and testing, respectively, by designating the 9^{th} utterances from both the training and development sets as our validation set. The remaining 8 utterances from the training and the development set constitute our primary training and testing set, respectively, for reporting experimental results.

Each utterance is annotated with continuous arousal and valence ratings within the range of $[-1, 1]$ the by 6 human raters in a sampling rate of 40ms. Following the suggestion in [13], the original ratings initially map to the range $[0, 1]$ using a linear transformation $y = 0.4975x + 0.5$ to conform to the requirements of Beta distributions. To increase the number of sample points for inferring target distributions and capture temporal information of the emotion perceptions, we concatenate the ratings from neighboring time frames. Subsequently, the target distributions are computed from the new label sets following the steps outlined in Section 2.1.

3.2. Simulating Noisy Condition

To replicate the signal in real world conditions, we added the noise signal to the original signal in testing set by setting the SNR level set:{0dB, 5dB, 10dB, 15dB, 20dB, 25dB} to mimic different noisy conditions. The 5-minute noise signal is obtained by concatenating the randomly selected noise audios from the noise part of the MUSAN corpus [17], which is a common dataset used to mimic the noisy condition in several speech emotion studies [10, 18]. Each utterance in different SNR levels is added by different noise audio combinations to ensure the noise diversity is sufficient.

3.3. Experimental Setup

Speech feature: The Bag-of-audio-words (BoAW) feature set is utilized in our experiment owing to its state-of-the-art performance in ambiguity modelling tasks [6]. From each utterance, we extract 39-dimensional MFCCs using a 25 ms window and a 10 ms hop size with torchaudio [19]. The audio codebook for BoAW is generated through k-means++ clustering, standardizing input for each 3-second window length, resulting in a 100-dimensional features set. The features are extracted using the openXBOW toolbox [20].

System configurations: Implementation is utilized by PyTorch 1.12.1 [21], with model parameters initialized using PyTorch's default settings. For enhanced training efficiency, training speech utterances are segmented into 3-second chunks, and the training batch size is capped at 100 due to GPU memory constraints. Models are optimized using an Adam optimizer with learning rate of 0.01. Decay ratios of 0.9, 0.99, and 0.999 are applied at the 10th, 20th, and 30th iterations, respectively. The maximum number of iterations is set at 100, with early stopping based on the best loss performance on the validation set.

Hyperparameters: Hyperparameter tuning is conducted through grid search, exploring values for distribution characteristic loss weights ($\eta_\mu \in [1, 20]$, $\eta_\sigma \in [1, 20]$), model dimensions in $32, 64, 128$, and $A \in \{5k | k \in [1, 6]\}$. The number of model parameters are from 25k to 250k, depended on model dimensions. The total training process spans approximately 48 hours on an Nvidia GeForce GTX 1060 6GB.

Evaluation metric: The distribution characteristics (μ and σ) is evaluated using the CCC as per literature [4], denoted as CCC_μ and CCC_σ, respectively. Belief Mismatch Coefficient (BMC) is employed for further evaluating the overall performance of the system when ambiguity is modelled. All experiments are tested under 7 noisy conditions, including 6 different levels of Signal-to-Noise Ratio (SNR) testing sets and the clean (original) testing set.

4. Experimental Results

4.1. Experiments Overview

- **Exp1: Impact of Modelling Ambiguity in Noisy Conditions**: This experiment aims to assess whether incorporating ambiguity modelling enhances the noise robustness of CER systems. We compare two systems: Vanilla vs. Ambiguity. The Vanilla system does not account for inter-rater ambiguity. It trains two separate models to independently predict the mean and SD of the target Beta distributions. In contrast, the Ambiguity model follows the system structure outlined in section 2, which simultaneously models μ and σ. It is optimized using the loss function in Equation (1) with $\eta_\gamma = 0$.
- **Exp2: Modeling With Belief Mismatch Loss (BML)**: This experiment aims to further evaluate the effectiveness of incorporating ambiguity into the system training through the proposed belief mismatch loss function. We test the model, denoted as Ambiguity$_{BML}$, which is trained using Equation (1), and compare its performance with both the Vanilla and Ambiguity models.

4.2. Result Discussion

Predicting valence from speech data is inherently challenging and tends to yield lower performance, even in clean conditions, compared to arousal [22]. For instance, the CCC_μ and CCC_σ values for the Vanilla model with clean speech data are 0.248 and 0.044, respectively, considerably lower than those for arousal as indicated in Table 1. Thus, to comprehensively analyze our hypothesis in noisy conditions, we focus exclusively on the arousal dimension. The best-performing hyperparameter set $(\eta_\mu, \eta_\sigma, H_{dim}, A)$ for reporting the CCC results are $(1, 14, 64, 5)$, respectively.

4.2.1. Exp1 Result Discussion

Figure 2 illustrates the CCC comparisons between the Vanilla and Ambiguity models. It is evident that the performance of the Ambiguity model, in terms of CCC_μ and CCC_σ, remains relatively stable as the SNR decreases, with an average degradation of (0.124, 0.050). In contrast, the performance of the Vanilla model degrades much faster, with an average degradation of (0.257, 0.158). The detailed CCC values are also reported in the first two rows of Table 1. These results strongly support the hypothesis that incorporating inter-rater ambiguity during training leads to greater noise robustness in CER systems compared

[1] https://github.com/crowpeter/BMLoss

Table 1: *The CCC results of distribution characteristics μ and σ for experiment 1 and 2 in each noisy condition.*

	CCC_μ								CCC_σ							
	clean	25dB	20dB	15dB	10dB	5dB	0dB	Avg.	clean	25dB	20dB	15dB	10dB	5dB	0dB	Avg.
Vanilla	0.661	0.582	0.513	0.436	0.316	0.320	0.257	0.441	**0.387**	0.317	**0.351**	0.208	0.192	0.188	0.119	0.275
Ambiguity	0.648	0.612	0.614	0.531	0.462	0.470	0.452	0.541	0.345	0.320	0.319	0.287	0.323	0.269	0.252	0.332
Ambiguity$_{BML}$	**0.662**	**0.632**	**0.628**	**0.552**	**0.494**	**0.476**	**0.459**	**0.558**	0.374	**0.350**	0.347	**0.303**	**0.342**	**0.291**	**0.256**	**0.352**

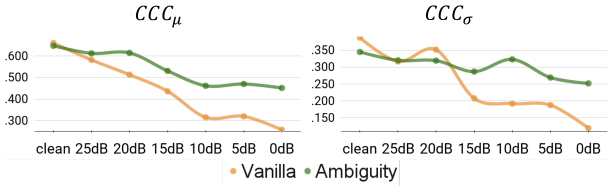

Figure 2: *Line chart visualization of CCC results for Vanilla and Ambiguity.*

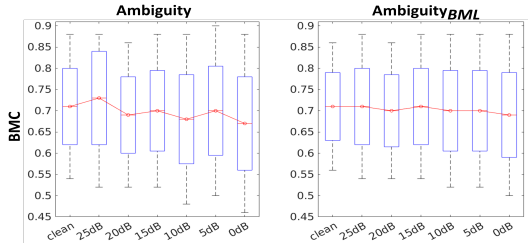

Figure 3: *BMC comparison across all conditions for Ambiguity and proposed Ambiguity$_{BML}$.*

to models that do not account for ambiguity. Additionally, on average, the Ambiguity model achieves a relative improvement of (0.100, 0.057) in terms of (CCC_μ, CCC_σ) compared to the Vanilla model, further highlighting the effectiveness of modelling ambiguity in enhancing noise robustness.

4.2.2. Exp2 Result Discussion

Figure 4 illustrates an example (dev4 utterance) of visualisation of the predicted distributions and the target distributions at an SNR level of 10dB. The shaded area indicates the $\mu \pm 1.5\sigma$ at each frame. It is obviously seen that the predicted distribution aligns significantly better with target distributions when ambiguity is modelled using Ambiguity$_{BML}$ while Vanilla loses ability to track emotions, further indicating the effectiveness of modelling ambiguity with proposed BML.

The comparison between Ambiguity and Ambiguity$_{BML}$ can be found in the 2^{nd} and 3^{rd} rows of Table 1. Ambiguity$_{BML}$ outperforms in both CCC_μ and CCC_σ across all conditions, demonstrating a relative improvement of (0.017, 0.020) on (CCC_μ, CCC_σ), respectively. The results indicate that learning the overall shape of distributions during training can benefit ambiguity-aware CER systems in both clean and noisy conditions.

We further validate the proposed approach using BMC evaluations that measure the overall mismatch between the distributions as depicted in Figure 3. Each box indicates the mode (red bar) and the 1_{st} and 3_{rd} quartiles of the BMCs evaluated across the entire testing data. Across all conditions, the original Ambiguity model exhibits greater sensitivity to noise, as evidenced by a larger variation in BMC compared to the Ambiguity$_{BML}$. These results significantly support that modelling ambiguity through the proposed BML enhance noise robustness.

4.3. Analysis: Noise Augmentation Validation

Based on the findings above that modelling ambiguity lead to noise robustness in CER, it is worth exploring a comparative analysis between the proposed Ambiguity$_{BML}$ model and the common strategy for building noise-robust system. In this section, we compare the two approaches under different noisy conditions: noise augmentation and proposed Ambiguity$_{BML}$ model. With the similar setting as we described in Section 3.2, the augmented noisy signals mimic with six levels of Signal-to-Noise Ratios (SNRs). In each training iteration, one randomly selected noisy utterance is augmented and jointly trained with the clean set.

Table 2: *The comparison of noisy augmentation methods.*

	Clean	25dB	20dB	15dB	10dB	5dB	0dB	Avg.
CCC_μ								
Vanilla+Aug	0.582	0.573	0.580	0.569	0.517	0.516	0.582	0.560
Amb.$_{BML}$	0.662	0.632	0.628	0.552	0.494	0.476	0.459	0.558
Amb.$_{BML}$+Aug	0.649	**0.635**	**0.645**	**0.620**	**0.605**	**0.542**	**0.587**	**0.612**
CCC_σ								
Vanilla+Aug	0.330	0.322	0.305	0.331	0.317	0.290	0.355	0.351
Amb.$_{BML}$	0.374	0.350	0.347	0.303	0.342	0.291	0.256	0.352
Amb.$_{BML}$+Aug	**0.375**	**0.386**	0.322	**0.393**	**0.398**	**0.321**	**0.414**	**0.403**

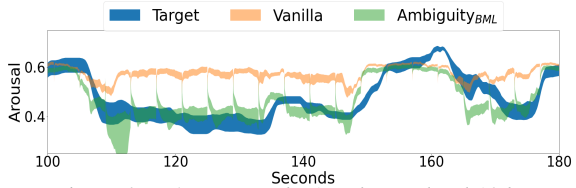

Figure 4: $\mu \pm 1.5\sigma$ *on dev1 with SNR level 10dB.*

As illustrated in Table 2, noise augmentation performs better in relatively noisier environments (SNR levels below 15dB), whereas Ambiguity$_{BML}$ achieves better performance in comparatively cleaner conditions (SNR levels above 20dB and clean). On average across all conditions, the performance of the two methods is closely aligned. This evidence demonstrates that modelling emotion perceptions from different raters, i.e., ambiguity, can yield a similar effect to the general noise augmentation method.

When applying noise augmentation to the proposed Ambiguity$_{BML}$, the results are further improved across most conditions, providing the feasibility of integrating ambiguity modelling with other approaches to further enhance noise robustness emotion recognition systems.

5. Conclusion and Future Work

In this study, we present a fresh perspective on modelling a noise-robust system in Continuous Emotion Recognition (CER) by incorporating inter-rater ambiguity. We introduce a novel Belief Mismatch Loss (BML) method, enabling the integration of ambiguity into system training. Experimental results reveal that ambiguity-aware CER consistently maintains stable performance across various noisy scenarios compared to systems lacking ambiguity modelling, particularly demonstrating superior performance when BML is employed during training. Interestingly, we observe that ambiguity modelling can yield comparable results to traditional noisy augmentation techniques in noisy environments, and with both approaches being complementary, combining them leads to performance improvements. As further analysis of valence performance in noisy environments was not pursued in this study, future work will focus on integrating inter-rater ambiguity on valence and signal variability caused by noise into the training process.

6. References

[1] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.

[2] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 373–380.

[3] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3507–3511.

[4] J. Wu, T. Dang, V. Sethu, and E. Ambikairajah, "A novel sequential monte carlo framework for predicting ambiguous emotion states," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8567–8571.

[5] M. Atcheson, V. Sethu, and J. Epps, "Using gaussian processes with lstm neural networks to predict continuous-time, dimensional emotion in ambiguous speech," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 718–724.

[6] D. Bose, "Continuous emotion prediction from speech: Modelling ambiguity in emotion," Ph.D. dissertation, UNSW Sydney, 2023.

[7] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards Robust Speech Emotion Recognition Using Deep Residual Networks for Speech Enhancement," in *Proc. Interspeech 2019*, 2019, pp. 1691–1695.

[8] S. Kshirsagar, A. Pendyala, and T. H. Falk, "Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions," *Frontiers in Computer Science*, vol. 5, p. 1039261, 2023.

[9] A. Wilf and E. M. Provost, "Towards noise robust speech emotion recognition using dynamic layer customization," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[10] Y.-T. Wu and C.-C. Lee, "MetricAug: A Distortion Metric-Lead Augmentation Strategy for Training Noise-Robust Speech Emotion Recognizer," in *Proc. INTERSPEECH 2023*, 2023, pp. 3587–3591.

[11] J. Wu, T. Dang, V. Sethu, and E. Ambikairajah, "Belief mismatch coefficient (bmc): A novel interpretable measure of prediction accuracy for ambiguous emotion states," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[12] D. Bose, V. Sethu, and E. Ambikairajah, "Parametric Distributions to Model Numerical Emotion Labels," in *Proc. Interspeech 2021*, 2021, pp. 4498–4502.

[13] ——, "Continuous emotion ambiguity prediction: Modeling with beta distributions," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–12, 2024.

[14] B. T. Atmaja and M. Akagi, "Evaluation of error-and correlation-based loss functions for multitask learning dimensional speech emotion recognition," in *Journal of Physics: Conference Series*, vol. 1896, no. 1. IOP Publishing, 2021, p. 012004.

[15] 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.

[16] 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.

[17] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[18] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6324–6328.

[19] Y. Y. Yang, M. Hira, Z. Ni, A. Astafurov, C. Chen, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang *et al.*, "Torchaudio: Building blocks for audio and speech processing," in *47th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2022*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 6982–6986.

[20] M. Schmitt and B. Schuller, "openxbow–introducing the passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[22] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Proc. Interspeech 2012*, 2012, pp. 1179–1182.