



Enrolment-based personalisation for improving individual-level fairness in speech emotion recognition

Andreas Triantafyllopoulos^{1,2}, Björn Schuller^{1,2,3,4}

¹CHI – Chair of Health Informatics, MRI, Technical University of Munich, Germany

²MCML – Munich Center for Machine Learning ³MDSI – Munich Data Science Institute

⁴GLAM – Group on Language, Audio, & Music, Imperial College London, UK

andreas.triantafyllopoulos@tum.de

Abstract

The expression of emotion is highly individualistic. However, contemporary speech emotion recognition (SER) systems typically rely on population-level models that adopt a ‘one-size-fits-all’ approach for predicting emotion. Moreover, standard evaluation practices measure performance also on the population level, thus failing to characterise how models work across different speakers. In the present contribution, we present a new method for capitalising on individual differences to adapt an SER model to each new speaker using a minimal set of enrolment utterances. In addition, we present novel evaluation schemes for measuring fairness across different speakers. Our findings show that aggregated evaluation metrics may obfuscate fairness issues on the individual-level, which are uncovered by our evaluation, and that our proposed method can improve performance both in aggregated and disaggregated terms.

Index Terms: personalisation, fairness, speech emotion recognition, computational paralinguistics, deep learning

1. Introduction

Speech emotion recognition (SER) was one of the earliest computational paralinguistic tasks to be tackled by machine learning (ML) and remains a core focal point for the speech community [1]. Progress in SER research has largely been measured in terms of gains in performance, with recent advances in deep learning (DL) spearheading current efforts [2]. This performance is usually measured in terms of a single metric computed over a held-out test set, typically accuracy or unweighted average recall (UAR) for classification or correlation for regression.

However, recent investigations on the *bias* of ML methods have called for increased attention to alternative performance evaluations that account for this bias. Related efforts have been largely geared towards identifying and mitigating bias resulting from certain *group* characteristics, such as biological sex, gender, or ethnicity [2, 3]. However, some works are calling attention to the need for an individual-level measure of fairness [2, 4], i. e., quantifying performance separately for each speaker.

This is related to the well-known problem of diverging performance across different speakers in several speech technology tasks (e. g., see Doddington’s zoo [5]), a topic which is becoming increasingly relevant when considering recent ethical and legal guidelines. According to the EU AI Act regulations adopted by the European Parliament [6, Amendment 52, “Proposal for a regulation”, Recital 26 c]:

There are serious concerns about the scientific basis of AI systems aiming to identify or infer emotions, particularly as expression of emotions vary considerably across cultures and situations, and even within a single individual. [...] Therefore, AI

systems identifying or inferring emotions or intentions of natural persons on the basis of their biometric data may lead to discriminatory outcomes and can be intrusive to the rights and freedoms of the concerned persons.

This quote illustrates the importance that lawmakers place on the generalisability of SER systems across different individuals, which in turn requires researchers to ensure equal treatment for all users of an SER system.

On a related note, there has been increased interest in improving the performance of SER using *personalisation*. This approach acknowledges that the expression of emotion can be highly individualistic [7]. Differences in expression may be influenced by culture [8], age [9], gender [10], or other factors, but also by temperamental differences across individuals [7, 11] Personalisation methods go beyond *population-level* models (i. e., models which are trained once and applied as-is to all new speakers during inference) and instead rely on *speaker-level* models which typically fall under three categories:

a) Methods which build speaker-level models; this can be done either by training/fine-tuning on data from a given individual, or training new models on subsets of the training data suited to that individual (e. g., similar speakers) [12, 13, 14, 15]. A typical example is that of Rudovic *et al.* [12], who introduce individual-level output heads; the data is first passed through a common backbone network trained for all speakers, and subsequently through output layers that are only trained with data from each speaker. This approach is similar to federated learning, which makes local updates to models using speaker data while simultaneously keeping a global model which aggregates model updates from all speakers [16]. Alternatively, a model can be retrained with data of the target speaker, or (additionally) with data from the most similar speakers to the target one during training [13]. The downside of these methods is that they require data from the target speaker to be already available during training, or on-the-fly retraining for each new speaker.

b) Methods which introduce additional personal information, e. g., in the form of demographic metadata (sex, age, etc.) that are given as extra inputs to the model [17, 18]. This approach is typically pursued in the scope of *precision medicine* [19], which aims to provide personalised treatment by accounting for individual patient histories and characteristics. While previous works have shown to improve performance with this approach on speech-based tasks as well [18], they do not capitalise on information about how a speaker actually sounds as they only exploit different ‘modalities’, thus leaving a gap to be covered by the last family of methods.

c) Methods which adapt population-level models in a *few-shot* fashion; typically, those rely on a few enrolment samples [4, 20, 21]. The upside of these approaches is that they require

no re-training of the model during inference. This allows for a streamlined deployment phase without any changes to the model – a significant benefit given the ever-increasing complexity of contemporary deep neural networks (DNNs). Examples include Triantafyllopoulos *et al.* [20] and Fan *et al.* [4], who use one or more neutral samples to condition an SER network, or Triantafyllopoulos *et al.* [21] who use two randomly selected samples in an analogous way. Similarly, Rahman and Busso [22] iteratively normalise features for each speaker; the added benefit of this method is that it does not require any label information, although more samples are required to obtain robust normalisation parameters and the method might not be suited to DNNs [20]. We note that these methods are different from attempts to extract speaker-specific emotion predictions by disentangling emotional and speaker information [23, 24]; essentially, personalisation, as we define it, boils down to *adaptation* to individual speaker characteristics, rather than *invariance* to them.

In the present contribution, we expand on both research directions. We introduce alternative considerations for *individual fairness*, inspired by definitions of *utility* and *fairness* in economic theory. Additionally, we revisit *personalisation via enrolment* in an attempt to improve on those metrics. Our methods are described in Section 2, with results and discussion following in Section 3. We summarise in Section 4.

2. Methodology

In this section, we begin with a description of the datasets we used in Section 2.1, followed by our personalisation method in Section 2.2, and experimental settings in Section 2.3. Our individual-level evaluation scheme is presented in Section 2.4.

2.1. Datasets

FAU-AIBO is a standard, categorical SER dataset used in the INTERSPEECH 2009 Emotion Challenge [25]. The data is collected in a Wizard-of-Oz scenario where a remotely-controlled robot (‘AIBO’) interacts with children between the ages of 6 and 10. The study participants attended one of two schools from the same region in Germany, (*Ohm*) and (*Mont*), with *Ohm* being used as the training set and *Mont* as the test set in the original challenge partition scheme. As no validation set was defined, we use the last two speakers (*Ohm*₃₁, *Ohm*₃₂) of the training set, similar to [20]. The collected data has been annotated for 11 classes by 5 individual raters on the word-level.

For the challenge, the 11 original classes were mapped to two alternative formulations of emotion: a 2-class problem, where participants had to differentiate between *negative* (*NEG*) and *non-negative* (*IDL*) emotions; and a 5-class problem, where participants had to classify an utterance as *angry* (*A*), *neutral* (*N*), *mothereseljoyful* (*P*), *emphatic* (*E*), with a 5th *rest* (*R*) class. Additionally, the original words were manually aggregated to semantically and prosodically meaningful chunks, with a chunk label derived from the word-level labels using a heuristic process [26]. The resulting data is highly imbalanced, and are dominated by neutral/non-negative states. For this reason, we use the UAR – the added recall per class divided by the number classes – to measure performance in the presence of class distribution imbalance, following the challenge specifications.

MSP-Podcast is a recent, large-scale SER dataset annotated both for categorical emotions and dimensional attributes [27]. The data has been annotated for 9 emotional classes, plus an extra neutral class and another one for instances where annotators disagree. We use the latest version available at the time of

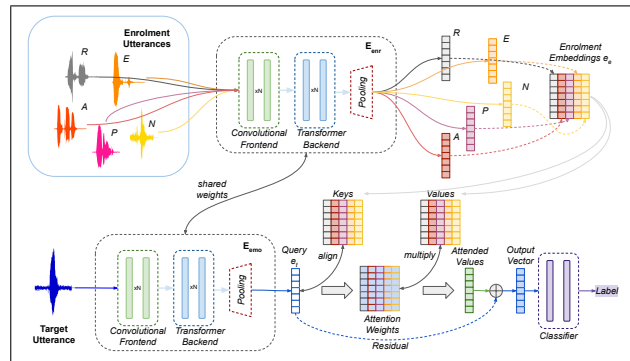


Figure 1: *Overview of the proposed architecture. A set of enrolment utterances is passed through an encoder to generate enrolment embeddings. These are used as keys and values in a dot-product attention scheme with the embeddings generated from the target utterance. The output embedding is passed to a feed-forward neural network for the final classification. Weights are shared between the enrolment and the main encoders.*

submission (v1.11.0), and focus exclusively on the standard 4-class problem pursued in most other recent SER works [28], with the set of labels being {*angry*, *happy*, *neutral*, *sad*} and exclude all other data. We end up with 44 586 instances in the training set, 11 947 in the validation, and 20 845 in the test set. We only use the more recent TEST1 partition and exclude TEST2.

2.2. Personalisation via enrolment

An overview of our architecture is shown in Fig. 1¹. Our workflow encapsulates the following key principles:

- a) Following standard DL practice, the target utterance u_t is passed through an emotion encoder $E_{emo}(\cdot)$ to generate a target embedding e_t . $E_{emo}(\cdot)$ can be any DNN; here, we opt for a WAV2VEC2.0 model fine-tuned for dimensional SER [2], as we expect it to already have a good representation of emotion.
- b) A set of *enrolment* utterances $\{u_e\}$, coming from the same speaker as the target utterance, are used to *adapt* an SER system to new, previously unseen speakers. In general, these $\{u_e\}$ can be exemplars of every potential label supported by the SER model or a subset thereof. In practice, we experiment with different alternatives as discussed below.
- c) These enrolment utterances are passed through an enrolment encoder $E_{enr}(\cdot)$ to generate suitable embeddings $\{e_e\}$. $E_{enr}(\cdot)$ can either be different from the $E_{emo}(\cdot)$ used for u_t , be identical to it but trained separately, or even share the same weights. We opt for the latter option as we intend the enrolment encoder to capture the emotional information present in $\{u_e\}$ and project it to the same embedding space as the target utterance.
- d) We then combine the enrolment embeddings $\{e_e\}$ with the target embedding e_t . We choose a multihead, dot-product attention mechanism for this combination. The enrolment embeddings are first concatenated and form a set of *keys* (K), with the target embedding functioning as the *query* (Q). Following Vaswani *et al.* [29], we first compute the outer product of the keys and queries (QK^T) and then pass it through a softmax function to generate a ‘soft’ similarity estimate of the query with each key; for numerical stability the output is divided by $\sqrt{(d_k)}$, with d being the dimensionality of the key embeddings. This soft similarity matrix is subsequently multiplied with the *values* (V)

¹Code: <https://github.com/ATriantafyllopoulos/enrollment-personalization>

to produce the final output which is added to the target embedding to generate the final output (i. e., the attention mechanism features a ‘residual’ connection). As values, we also use the embeddings of the enrolment utterances.

e) The final product of the attention is passed to a multi layer perceptron (MLP) classifier and the whole system is trained end-to-end.

Intuitively, this process of attention computes a (soft) similarity of the target utterance with each enrolment utterance, then combines the emotional information in these enrolment samples with the target sample to improve classification performance.

2.3. Experimental settings

Enrolment utterances² Our approach requires setting aside a set of enrolment utterances for each speaker. We always include a single enrolment utterance per class (though some variants of our approach do not make use of all of them; see below). When a class is missing for a particular speaker, we impute it with zeros (i. e., silence). To create this enrolment set, we first sort the utterances of each speaker alphabetically (so in their order of appearance following the naming scheme of FAU-AIBO [26] and MSP-Podcast [27]) and then for each class select the first utterance in which it appears.

Personalisation setup: In total, we investigate three alternative formulations of the personalisation problem and contrast their effectiveness with respect to a baseline model:

1. **BASE** – Our baseline model only includes the emotion encoder and the downstream MLP classifier; note that this setup is identical to recent state-of-the-art work for dimensional SER [2] and we thus expect it to be a strong baseline.
2. **PERS_N** – For our first personalisation approach, we only utilise neutral utterances for the enrolment; this is most similar to the setup of Triantafyllopoulos *et al.* [20], who condition their model on a single neutral utterance from each speaker.
3. **PERS_E** – We also experiment with using only the non-neutral (i. e., emotional or rest) utterances for enrolment.
4. **PERS_A** – Finally, we use all available enrolment utterances (both neutral and emotional).

Hyperparameters: We train all models for 50 epochs with an Adam optimiser, a learning rate of 0.0001, and a batch size of 4, all standard hyperparameters from previous literature [2]. We select the epoch with the best UAR on the validation set for our final evaluation on the test set.

2.4. Individual-level fairness

The standard process for computing performance is to consider each chunk as an independent trial. This process was also adopted in the 2009 INTERSPEECH Emotion Challenge, and we denote its outcome as UAR_C .

To define individual-level fairness, we begin by computing the performance on a speaker-level, thus assuming that first speakers are selected independently, and only then are samples selected independently for them (see [30] for a similar argumentation). We finally compute the UAR over the set of chunks for each individual speaker in our dataset, which we denote as UAR_{SP} . We call this the **utility** of each speaker, as this is the benefit that each speaker can expect from getting their emo-

tions recognised correctly³. We examine this utility under three different perspectives:

I) Statistics: We first report standard statistics, such as the mean ($\mu(\cdot)$), median ($\mu_{1/2}(\cdot)$), standard deviation ($\sigma(\cdot)$), maximum ($\max(\cdot)$), and minimum ($\min(\cdot)$) of UAR_{SP} . These statistics give us a coarse characterisation of how utility is distributed across the different individuals in our dataset.

II) Gini coefficient: The Gini coefficient ($G(\cdot)$) is a standard measurement used in econometrics to judge the distribution of utility in a particular population and is defined as half of the mean absolute difference relative to the mean of a particular sample [31]. In particular, it takes the value of 0 for an equal distribution where everyone has the same utility, and 1 for a completely unequal one, where the entire utility is accumulated by one particular individual. In our case, this is computed as:

$$G(\{u_i\}_1^N) = \frac{\sum_{i=1}^{i=N} \left(\sum_{j=1}^{j=N} (|u_i - u_j|) \right)}{\mu(\{u_i\}_1^N)}, \quad (1)$$

where N is the number of speakers in the test set and $\{u_i\}_1^N$ is the set of utility values for all speakers, with $u_i = UAR_{SP}^i$, i. e., the speaker-level UAR computed for each speaker.

III) Isoelastic social welfare functions (ISWFs): Besides the Gini coefficient, the distribution of utility can also be considered under alternative formulations. A classical formalisation of the problem is that given by Atkinson *et al.* [32], who defines a family of ISWFs:

$$W_\alpha(u_1, \dots, u_N) = \begin{cases} \frac{1}{N} \left(\sum_{i=1}^{i=N} u_i^{1-\alpha} \right)^{\frac{1}{1-\alpha}} & \text{if } \alpha \neq 1 \\ \sqrt[N]{\prod_{i=1}^{i=N} u_i} & \text{if } \alpha = 1, \end{cases} \quad (2)$$

These ISWFs allow for a more modular definition of fairness, as they do not presuppose an equal distribution of utility as the most fair outcome (like the Gini coefficient). For example, as $\alpha \rightarrow \infty$, Eq. (2) approximates Rawls’ ‘difference principle’ (i. e., maximise the minimum utility) [33, Ch. II, § 13, pg. 65], whereas as $\alpha \rightarrow 0$, it approaches the standard utilitarian approach of maximising total utility irrespective of fairness. Thus, ISWFs provide a modular ‘knob’ that allows stakeholders to define fairness for their particular needs. In our work, we compute the value of the ISWF for different values of $\alpha \in \{0, 100\}$, measuring the suitability of different models for different scenarios.

3. Results & Discussion

Table 1 shows our results. It includes the global UAR (UAR_C), which is computed over all chunks in the data (including 95% CIs), as well as the different fairness statistics and the Gini index for both the 2- and the 5-class problem. **PERS_A** shows the best performance overall, with higher UAR_C than all alternatives for FAU-AIBO and marginally lower than the baseline for MSP-Podcast, as well as overall better performance with respect to different individual-level metrics. Specifically, it improves from a 67.7% to a 71.8% UAR_C for the 2-class problem, and from 44.3% to 45.2% for the 5-class problem for FAU-AIBO. More importantly, it improves across all fairness metrics for all datasets and tasks, showing a lower Gini index and standard

²We include CSV files with the filename used for training/development/test set and the corresponding enrolment sets as supplementary material.

³Naturally, what this ‘benefit’ entails depends entirely on the downstream application.

Table 1: Global and individual-level performance for the 2- and 5-class problems of FAU-AIBO, as well as the 4-class problem of MSP-Podcast. We compute the UAR_C over all chunks in the test set, along with 95% confidence intervals (CIs) obtained via bootstrapping. Furthermore, we compute fairness metrics on speaker-level UAR (UAR_{SP}). For easier comparison, we mark metrics as ascending (\uparrow , higher is better) and descending (\downarrow , lower is better).

Method	$UAR_C(\uparrow)$	$G(UAR_{SP})(\downarrow)$	$\mu(UAR_{SP})(\uparrow)$	$\sigma(UAR_{SP})(\downarrow)$	$\mu_{1/2}(UAR_{SP})(\uparrow)$	$max(UAR_{SP})(\uparrow)$	$min(UAR_{SP})(\uparrow)$
FAU-AIBO (2-class problem)							
BASE	.677 [.666 - .689]	.091	.699	.115	.686	1.000	.500
PERS_N	.680 [.669 - .692]	.097	.697	.120	.715	1.000	.500
PERS_E	.700 [.690 - .711]	.082	.703	.107	.731	1.000	.500
PERS_A	.718 [.707 - .729]	.073	.728	.098	.735	1.000	.500
FAU-AIBO (5-class problem)							
BASE	.443 [.426 - .461]	.160	.406	.132	.417	.806	.053
PERS_N	.424 [.408 - .442]	.118	.399	.086	.390	.614	.214
PERS_E	.440 [.422 - .460]	.137	.432	.137	.419	1.000	.207
PERS_A	.452 [.434 - .469]	.106	.428	.084	.412	.576	.189
MSP-Podcast (4-class problem)							
BASE	.567 [.559 - .575]	.242	.438	.200	.413	1.000	.000
PERS_N	.510 [.502 - .520]	.233	.361	.170	.333	1.000	.000
PERS_E	.563 [.554 - .573]	.215	.460	.184	.436	1.000	.000
PERS_A	.563 [.554 - .572]	.201	.479	.186	.442	1.000	.000

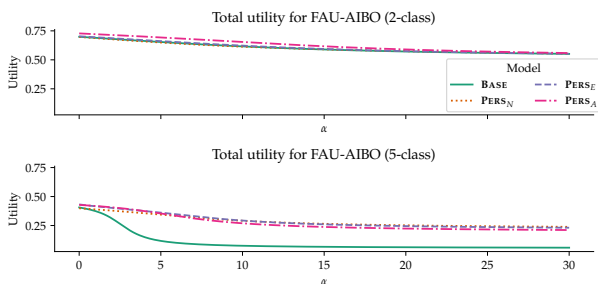


Figure 2: Total utility achieved by each model for different isoelastic social welfare functions for the 2- (top) and 5-class (bottom) formulations of FAU-AIBO. Utility is defined as UAR_{SP} .

deviation, as well as a higher mean and median for speaker-level UAR (UAR_{SP}) – our measure of utility. It is only surpassed by **PERS_N** on the standard deviation of performance for MSP-Podcast, where it is still outperforming the baseline model. Generally, the same can be said for **PERS_N** and **PERS_E**, as both show better performance than the baseline, with the latter additionally outperforming the former.

Fig. 2 additionally shows how the total utility amassed by each model changes for different ISWFs. Due to space limitations, we only show results for FAU-AIBO. For the 2-class problem, **PERS_A** results in higher utility for most values of α , while converging to the other methods as $\alpha \rightarrow \infty$. For the 5-class problem, we observe that **BASE** quickly approaches zero utility according to the difference principle ($\alpha \rightarrow \infty$), with the three personalisation models showing similar behaviour. For all models, utility is highest for $\alpha \rightarrow 0$, i. e., the utilitarian scenario, where average utility is maximised without consideration for its distribution. However, the total utility quickly drops as α increases, reflecting scenarios where the discrepancy between different speakers becomes more important.

Broadly, the decision on which the α value or fairness metric is appropriate for a particular application rests with the stakeholders that are affected by it. For example, recent work employed

an ‘equality of outcome’ requirement for different users across multiple ML algorithms (in the context of recommendation algorithms) [34]. This would be equivalent to a Gini index of 0 in our definition. Digital health applications, on the other hand, may require a maximisation of a lower bound on speaker-level performance – in this case, a larger α would be more appropriate, to place more emphasis on the worst-performing speakers [35].

Collectively, our results show that personalisation via enrolment can improve predictive performance and make this performance more equal across different speakers. This is vital for providing a uniform and fair user experience in applications relying on SER. Another interesting finding is that models personalised on all classes or even only all emotional classes generally outperform those personalised only on neutral data. This is in contrast to the prior work which has used these neutral enrolment utterances for personalisation [4, 20].

4. Conclusion

We have introduced a novel method for personalisation using a minimal set of enrolment utterances – one per class. Our method relies on dot-product attention for injecting information from these utterances into a classification network. Additionally, we introduced novel considerations for individual-level fairness that takes into account performance on the individual level. Overall, we showed how our method can improve performance both on the global level and for the fairness metrics we introduced. **Limitations:** Our method depends on providing accurate enrolment samples, and may fail if (intentionally) given erroneous ones. **Future work:** Alternative methods for introducing the enrolment information can be investigated. Furthermore, explainability methods can be used to understand how the network is using the additional enrolment information, e. g., by visualising and probing the embedding space before and after the enrolment information is injected. Finally, ways for safeguarding against inaccurate enrolment samples must be developed.

5. Acknowledgements

This work was partially funded by the EU H2020 project No. 101135556 (INDUX-R).

6. References

- [1] B. 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.
- [2] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 09, pp. 10 745–10 759, 2023.
- [3] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," in *Interspeech*, 2019, pp. 2823–2827.
- [4] W. Fan, X. Xu, B. Cai, and X. Xing, "Isnet: Individual standardization network for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1803–1814, 2022.
- [5] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *Proc. ICSLP*, Sydney, Australia: ISCA, 1998, pp. 1–4.
- [6] The European Parliament, *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))*, <https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236.EN.html>, 2023.
- [7] R. J. Larsen and E. Diener, "Affect intensity as an individual difference characteristic: A review," *Journal of Research in Personality*, vol. 21, no. 1, pp. 1–39, 1987.
- [8] J. A. Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies," *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.
- [9] J. J. Gross, L. L. Carstensen, M. Pasupathi, J. Tsai, C. Götestam Skorpen, and A. Y. Hsu, "Emotion and aging: Experience, expression, and control," *Psychology and aging*, vol. 12, no. 4, p. 590, 1997.
- [10] T. M. Chaplin and A. Aldao, "Gender differences in emotion expression in children: A meta-analytic review," *Psychological bulletin*, vol. 139, no. 4, p. 735, 2013.
- [11] R. A. Sherman, J. F. Rauthmann, N. A. Brown, D. G. Serfass, and A. B. Jones, "The independent effects of personality and situations on real-time expressions of behavior and emotion," *Journal of personality and social psychology*, vol. 109, no. 5, p. 872, 2015.
- [12] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, pp. 1–11, 2018.
- [13] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1959–1972, 2022.
- [14] M. Song, A. Triantafyllopoulos, Z. Yang, H. Takeuchi, T. Nakamura, A. Kishi, T. Ishizawa, K. Yoshiuchi, X. Jing, V. Karas, et al., "Daily mental health monitoring from speech: A real-world japanese dataset and multitask learning analysis," in *Proc. ICASSP*, IEEE, 2023, pp. 1–5.
- [15] A. Kathan, M. Harrer, L. Küster, A. Triantafyllopoulos, X. He, M. Milling, M. Gerczuk, T. Yan, S. T. Rajamani, E. Heber, et al., "Personalised depression forecasting using mobile sensor data and ecological momentary assessment," *Frontiers in digital health*, vol. 4, p. 964 582, 2022.
- [16] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.
- [17] T. Hulsen, S. S. Jamuar, A. R. Moody, J. H. Karnes, O. Varga, S. Hedensted, R. Spreafico, D. A. Hafler, and E. F. McKinney, "From big data to precision medicine," *Frontiers in medicine*, vol. 6, p. 34, 2019.
- [18] M. Gerczuk, A. Triantafyllopoulos, S. Amiriparian, A. Kathan, J. Bauer, M. Berking, and B. W. Schuller, "Zero-shot personalization of speech foundation models for depressed mood monitoring," *Patterns*, vol. 4, no. 11, 2023.
- [19] M. R. Kosorok and E. B. Laber, "Precision medicine," *Annual review of statistics and its application*, vol. 6, pp. 263–286, 2019.
- [20] A. Triantafyllopoulos, S. Liu, and B. W. Schuller, "Deep speaker conditioning for speech emotion recognition," in *Proc. ICME*, IEEE, 2021, pp. 1–6.
- [21] A. Triantafyllopoulos, M. Song, Z. Yang, X. Jing, and B. W. Schuller, "Exploring speaker enrolment for few-shot personalisation in emotional vocalisation prediction," *arXiv preprint arXiv:2206.06680*, 2022.
- [22] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2012, pp. 5117–5120.
- [23] C. Le Moine, N. Obin, and A. Roebel, "Speaker attentive speech emotion recognition," in *Interspeech 2021*, ISCA, 2021, pp. 2866–2870.
- [24] Y. Yin, B. Huang, Y. Wu, and M. Soleymani, "Speaker-invariant adversarial domain adaptation for emotion recognition," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 481–490.
- [25] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proc. INTERSPEECH*, ISCA, Brighton, UK: ISCA, Sep. 2009, pp. 312–315.
- [26] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin: Logos Verlag, 2009, (PhD thesis, FAU Erlangen-Nuremberg).
- [27] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.
- [28] A. Triantafyllopoulos, U. Reichel, S. Liu, S. Huber, F. Eyben, and B. W. Schuller, "Multistage linguistic conditioning of convolutional layers for speech emotion recognition," *Frontiers in Computer Science*, vol. 5, 2023.
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [30] I. Guyon, J. Markhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error rate estimates?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 52–64, 1998.
- [31] R. Dorfman, "A formula for the gini coefficient," *The review of economics and statistics*, pp. 146–149, 1979.
- [32] A. B. Atkinson et al., "On the measurement of inequality," *Journal of economic theory*, vol. 2, no. 3, pp. 244–263, 1970.
- [33] J. Rawls, *A theory of justice: Revised Edition*. Harvard University Press, 2009.
- [34] S. Sharifi-Malvajerdi, M. Kearns, and A. Roth, "Average individual fairness: Algorithms, generalization and experiments," *Proc. NeurIPS*, vol. 32, 2019.
- [35] A. Triantafyllopoulos, A. Kathan, A. Baird, L. Christ, A. Gebhard, M. Gerczuk, V. Karas, T. Hübner, X. Jing, S. Liu, et al., "Hear4health: A blueprint for making computer audition a staple of modern healthcare," *Frontiers in Digital Health*, vol. 5, p. 1 196 079, 2023.