Testing Speech Emotion Recognition Machine Learning Models

Anna Derington, Hagen Wierstorf, Ali Özkil, Florian Eyben, Felix Burkhardt, Björn W. Schuller

***Abstract*—Machine learning models for speech emotion recognition (SER) can be trained for different tasks and are usually evaluated on the basis of a few available datasets per task. Tasks could include arousal, valence, dominance, emotional categories, or tone of voice. Those models are mainly evaluated in terms of correlation or recall, and always show some errors in their predictions. The errors manifest themselves in model behaviour, which can be very different along different dimensions even if the same recall or correlation is achieved by the model. This paper investigates behavior of speech emotion recognition models with a testing framework which requires models to fulfill conditions in terms of correctness, fairness, and robustness.**

I. Introduction

Machine learning models are developed to fulfill a given objective by presenting them examples. In speech emotion recognition (SER) we might have the prediction of arousal as objective and audio samples with an associated arousal value as examples. Development pipelines are restricted to certain model specifications and a limited amount of examples leading to non-perfect fulfillment of the objectives. This requires further evaluation steps to estimate the performance of the models. The evaluation can focus on tracking progress in a given field by specifying benchmarks in order to compare models [1]. As much of the progress in recent times is achieved by general-purpose foundation models [2] the focus has shifted towards more and smaller datasets as benchmarks [3] and a larger variety of tasks [4].

The inductive biases of trained models can lead to models containing spurious correlations or learnt shortcuts due to the underspecification of the applied development pipeline [5]. This means two models showing the same accuracy performance in a benchmark might have very different general behaviour and properties, that need to be understood [2, §4.4] and communicated to stakeholders [6].

As it is most often required that a model stays in a certain range of expected behaviour, testing is a valid evaluation approach as it can detect differences between existing and required behaviour [7]. Testing machine learning models provides a greater challenge compared to software testing [8] as the models provide answers to questions for which no previous answer – e. g., label – was available (oracle problem) [9]. This can be solved by changing available input samples from test sets in a way that labels are preserved [10] or with an expected change of labels [11]. In addition, synthetic data with known labels might be generated [12].

In SER no general testing approach was proposed so far. The Computational Paralinguistics Challenges [13] tracked progress for SER and have led to rapid progress in the area. Scheidwasser-Clow *et al.* [14] have introduced a multi-dataset benchmark focusing on the evaluation of fine-tuned foundation models. Further, Jaiswal and Provost [15] have evaluated the robustness of a model that predicts categorical emotions under different data augmentations. Their model under test showed significant performance degradation for most of the applied augmentations. They also showed that some of the augmentations like change in pitch, adding laughter, crying, or speeding up the utterance can affect the underlying label of the human perceived emotion as well. Triantafyllopoulos *et al.* [16] presented a framework to estimate how much the prediction of valence of a SER model depends on the extracted sentiment from text instead on the tone of voice. Schmitz *et al.* [17] addressed the topic of fairness [18], which is otherwise rarely addressed in the SER community so far.

In this paper, we follow Zhang *et al.* [7] and propose to implement the evaluation of model behaviour in the form of offline tests. We focus on SER models with the regression task of predicting emotional dimensions (arousal, dominance, valence) and the classification task of predicting emotional categories (happy, anger, ...). The development of the tests is not driven by the concept of unit tests or coverage, but it is motivated by possible applications of the model, and how it should behave under certain conditions. We solve the test oracle problem for test data without labels by comparing predictions across protected groups, e. g., assuming that we should get the same distribution of model predictions for groups that differ by a certain aspect like speaker accent; or by generating different versions of labelled data with data augmentations for which we know they should not affect the label. The resulting testing suite is available at https://audeering.github.io/ser-tests/.

II. Method

Zhang *et al.* [7] group testing properties for machine learning systems into the categories correctness, model relevance, robustness, security, data privacy, efficiency, fairness, and interpretability. In this work, we focus on the properties **correctness**, **robustness**, and **fairness**. We do not cover model relevance, i. e., whether the complexity of the model fits the

e-mail: aderington@audeering.com

A. Derington, H. Wierstorf, F. Eyben, F. Burkhardt B. W. Schuller are with audEERING GmbH, Gilching, Germany.

A. Özkil is with Jabra, GN Audio, Copenhagen, Denmark

B. W. Schuller is with audEERING GmbH, Gilching, Germany; CHI – Chair of Health Informatics, Technical University of Munich, Germany; GLAM – Group on Language, Audio, & Music, Imperial College London, UK.

data, interpretability, and data privacy, as they are most commonly covered by means of white-box and grey-box testing and may be architecture specific, whereas we require our framework to include only universally applicable black-box tests. Security testing is often related to adversarial robustness, especially for tasks such as autonomous driving where facing adversarial examples introduces high risks [12]. For the task of speech emotion recognition, we assume that the model is only applied in scenarios with no or low risk and do not cover security in addition to robustness. We do not address efficiency with a dedicated set of tests, but the results of all tests can be used to compare smaller variants of a model [19] or models trained on a subset of the training data [20] of its original version.

For each test we propose a selection of evaluation metrics, and suggest a threshold that determines a passing or failing result. We propose a method to set the fairness test thresholds automatically based on numeric simulations and independent of the application. Whereas the thresholds for correctness and robustness need to be defined with certain applications in mind. When comparing models, it is also recommended to analyse the specific metric results, since a small change in model behaviour could be the difference between a passing and failing test.

A. Datasets

For our tests, we include a multitude of emotional datasets from various domains. In order to compare results between datasets, we map the categorical labels to a standardised set of names (e. g., anger, happiness, neutral, or sadness). We map samples labelled as joy to the category of happiness. On the dimensional labels we apply min-max scaling to scale them to the range of $[0, 1]$.

The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [21] is a dataset containing 7,442 samples from 91 actors (48 female, 43 male) between the ages of 20 and 74 years from various races and ethnicities. The actors were tasked to portray a selection of 12 English sentences in different emotions (anger, disgust, fear, happiness, neutral, and sadness) in varying levels of intensity. Emotion ratings are available for the audio modality, the visual modality, and for the combined modalities. We use the ratings from the audio modality, removing samples with no agreement from the dataset. As there is no official test set, we define our own split of the dataset with the samples from 17 speakers (8 female, 9 male) [22].

Danish Emotional Speech (DES) [23] is an approximately 30 minutes long dataset containing recordings in Danish from 4 actors (2 female, 2 male) who convey the emotions anger, happiness, neutral, sadness, and surprise. We use the entire dataset as the test set.

Berlin Database of Emotional Speech (EmoDB) [24] is a German dataset in which 10 actors (5 female, 5 male) each portray 10 sentences in the emotions anger, boredom, disgust, fear, happiness, neutral, and sadness. Recordings took place in an anechoic chamber. We select 2 female and 2 male speakers for our test split [22].

Italian Emotional Speech Database (EMOVO) [25] is an Italian dataset with recordings of 6 actors (3 female, 3 male) tasked to portray 14 sentences in the emotional states anger, disgust, fear, joy, neutral, sadness, and surprise. We use the entire dataset for our tests.

Interactive Emotional Dyadic Motion Capture (IEMOCAP) [26] is a corpus collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). 10 actors were recorded in dyadic sessions which include a scripted portion and an improvised portion, designed to elicit emotional data, resulting in approximately 12 hours of speech. Each of the 5 recorded sessions contains a male and a female speaker. The dataset has been annotated for the arousal, dominance, and valence dimensions as well as for categories (anger, disgust, excitement, fear, frustration, happiness, neutral, other, and sadness). We form a test split of the dataset from sessions 4 and 5, but use all sessions for certain tests where we require a higher number of speakers to gain relevant results [22].

Multimodal EmotionLines Dataset (MELD) [27] is an enhancement on the EmotionLines dataset [28] that extends it to include audio and visual modalities. It contains about 13,000 utterances from the English TV-series *Friends*, and has been annotated with emotion and sentiment labels. The emotion categories are anger, disgust, fear, joy, neutral, sadness, and surprise. We use the official test set, but removed files shorter than $0.76\,\mathrm{s}$ or longer than $30\,\mathrm{s}$ [22].

MSP-Podcast [29] is a large speech emotional dataset built from segments from English podcast recordings. The dataset is annotated using crowdsourcing for the arousal, dominance, and valence dimensions, and categorical labels (anger, contempt, disgust, fear, happiness, neutral, other, sadness, and surprise). We use version 1.7 of the dataset, which has roughly 100 hours of speech data. We evaluate our tests on both of the test sets: test set 1 with 30 male and 30 female speakers and test set 2 with approximately 3,500 segments from 100 podcasts not used in other partitions.

Polish Emotional Speech Database (PESD) [30] comprises 240 recordings in Polish from 8 actors (4 female, 4 male). Each speaker utters 5 sentences with 6 types of emotional prompts: anger, boredom, fear, joy, neutral, and sadness. We use the combined set of samples as a test set.

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [31] contains audio and video recordings from 24 professional actors (12 female, 12 male) vocalising 2 English statements in speech and song. Speech samples are expressed in the emotions anger, calm, disgust, fear, happiness, neutral, sadness, and surprise. Each emotion except for the neutral expression is produced in a normal as well as a strong intensity. For our test set, we select 2 female and 2 male speakers and only use their speech samples [22].

Some tests might use additional datasets, e. g., to add background noise. In those cases, the dataset is introduced in the test definition.

B. Correctness Tests

The correctness tests require that the model predictions follow the true labels as closely as possible. Correctness is

TABLE I
OVERVIEW OF THE CORRECTNESS TESTS, THEIR TEST SETS, METRICS, AND PASSING CONDITIONS.

Test	Task	Test Sets	Metric	Condition
Correctness Classification	categories	CREMA-D, DES, EmoDB, EMOVO, IEMOCAP, MELD, MSP-Podcast (test-1), MSP-Podcast (test-2), PESD, RAVDESS	precision per class (PPC)	> 0.5
			recall per class (RPC)	> 0.5
			unweighted average precision (UAP)	> 0.5
			unweighted average recall (UAR)	> 0.5
Correctness Consistency	dimensions	CREMA-D, DES, EmoDB, EMOVO, IEMOCAP, MELD, PESD, RAVDESS	Samples in Expected Range	> 0.75
Correctness Distribution	categories	CREMA-D, DES, EmoDB, EMOVO, IEMOCAP, MELD, MSP-Podcast (test-1), MSP-Podcast (test-2), PESD, RAVDESS	Relative Diff. Per Class	$\lvert\cdot\rvert < 0.15$
	dimensions	IEMOCAP, MSP-Podcast (test-1), MSP-Podcast (test-2)	Diff. Mean Value	$\lvert\cdot\rvert < 0.03$
			Jensen-Shannon Distance	< 0.2
Correctness Regression	dimensions	IEMOCAP, MSP-Podcast (test-1), MSP-Podcast (test-2)	concordance correlation coefficient (CCC)	> 0.5
			mean absolute error (MAE)	< 0.1
			Pearson correlation coefficient (PCC)	> 0.5
Correctness Speaker Average	categories	IEMOCAP, MELD, MSP-Podcast (test-1)	Class Proportion MAE	< 0.1
			Class Proportion mean directional error (MDE)	$\lvert\cdot\rvert < 0.05$
	dimensions	IEMOCAP (full), MSP-Podcast (test-1), MSP-Podcast (test-2)	MAE	< 0.1
			MDE	$\lvert\cdot\rvert < 0.05$
Correctness Speaker Ranking	categories	MELD, MSP-Podcast (test-1)	Precision Bottom 25%	> 0.7
			Precision Top 25%	> 0.7
			Spearmans Rho	$\lvert\cdot\rvert > 0.7$
			Top-Bottom Confusions	< 0.15
	dimensions	MSP-Podcast (test-1), MSP-Podcast (test-2)	Precision Bottom 25%	> 0.7
			Precision Top 25%	> 0.7
			Spearmans Rho	$\lvert\cdot\rvert > 0.7$
			Top-Bottom Confusions	< 0.15

a fundamental goal of most machine learning systems and as such rarely overlooked. However, special care is required to distinguish between different types of errors, some of which may matter more to a user than others. By looking at correctness from different viewpoints and with different metrics, more nuanced insights in model behaviour can be gained.

Table I shows a summary of the discussed correctness tests. Furthermore, it lists the test sets we apply the test metrics on, as well as the passing conditions we apply for each test metric. The thresholds for the passing conditions provide only an example to present our testing framework and to show how the test results can be summarized into a percentage of passed tests. We suggest to adjust the thresholds to the needs of each individual application, or to compare the metric results directly without enforcing a binary result of passing or failing. An alternative option would be to base the thresholds for correctness on average human rater performance, e. g. by randomly selecting a rater for each sample, or by averaging all individual raters' performance. The number of correctness tests are 66 for arousal, 72 for dominance, 76 for valence, and 196 for categorical emotions. In the following, we discuss correctness tests that need additional information to what is presented in Table I.

The **Correctness Consistency** tests check whether the models' predictions on dimensional tasks are consistent with the expected result for samples with certain categorical labels. For example, happiness is characterised by high valence and fear tends to coincide with low dominance. Based on comparing various literature results [32, 33, 34, 35], we expect a correspondence between dimensional values and emotional categories as presented in Table II and Figure 5, where dimensional values ≥ 0.55 are counted to be in the high range, values between 0.3 and 0.6 in the neutral range, and values ≤ 0.45 in the low range.

TABLE II
CORRESPONDENCE BETWEEN EMOTIONAL CATEGORIES AND DIMENSIONAL VALUES BASED ON LITERATURE REVIEW.

emotion	valence	arousal	dominance
anger	low	high	high
boredom	neutral	low	
disgust	low		
fear	low	high	low
frustration	low		
happiness	high		neutral
neutral	neutral	neutral	neutral
sadness	low	low	low
surprise		high	neutral

When calculating the Jensen-Shannon divergence [36] for the **Correctness Distribution** test, we bin the distributions into 10 bins.

Certain applications of SER models may be interested in the average emotional value for each speaker. The **Correctness Speaker Average** tests check whether the models' estimate of the average speaker value is close to the truth. For regression, we measure the mean absolute error (MAE) and mean directional error (MDE) between the true speaker average and

the predicted speaker average. For classification, as we cannot compute a single average value, we compare the proportions of samples that are assigned a certain class per speaker. We then calculate the MAE and the MDE in the estimated proportion of samples for each class. We only consider speakers with at least 10 samples for regression, and with at least 8 samples per class for classification. We apply the tests on test sets for which six or more speakers remain that fit the criteria.

We test a potential ranking of speakers based on their average, for instance to spot outliers on either side of the ranking. This is covered in the **Correctness Speaker Ranking** tests. The average values (as computed in the Correctness Speaker Average tests) are used to create a ranking. For classification, we create a separate ranking for each class label, e. g., for the anger class, we rank speakers with a higher proportion of anger samples higher, and speakers with a lower proportion of anger samples lower. The top and bottom of a ranking are often of particular interest, which we test by computing the precision of the upper and lower quantiles of the ranking (bottom 25% and top 25%). Finally, we consider Top Bottom Confusions, i. e., speakers that are ranked into the top (or bottom) quantile although they are in the opposite quantile in the true ranking.

C. Fairness Tests

Despite a long history of research in the field, a universal definition for fairness has not been established, neither in a general sense, nor when applied to machine learning [37]. Many widely used definitions of fairness state that no bias should exist for certain protected attributes [38]. In Mehrabi *et al.* [37], algorithmic fairness is grouped into three main types: *individual fairness*, which aims to give similar predictions to similar individuals, *group fairness*, which tries to treat different protected groups equally, and *subgroup fairness*, which combines the two previous approaches by selecting a group fairness constraint and checks whether the constraint applies across sets of combinations of protected attribute values. Individual fairness would require data with similar samples that differ only in the protected attribute, and subgroup fairness would require test sets with annotations for all types of fairness groups (e. g., accent or language) at the same time, both of which are not easily available. Thus, we employ different types of group fairness tests and distinguish between cases where the ground truth emotion label is known and where it is not. Note that the fairness tests should not be interpreted as proof that a model is fair when it passes the tests, but rather as an indicator that the model is likely not fair towards a certain protected group when it fails the tests. The tests in the following do not cover all relevant groups for which fairness should be considered, but they provide a start to be expanded on.

Statistical parity (or *demographic parity*) is a group fairness criterion which enforces that the model function $f(X, S)$, given the input data X and the protected group S, is statistically independent of S [39]. For the classification of classes $c \in C$, this is given when for all $s \in S$

$$\mathbb{P}(f(X,S) = c) = \mathbb{P}(f(X,S) = c | S = s). \tag{1}$$

We apply this to our tests for unlabelled data by comparing the class wise distributions of samples for the different group members, and require that the differences are below a given threshold. We refer to this metric as the relative difference per class. For a regression model with $f(X, S) \in [0, 1]$, statistical independence requires that for all $s \in S$ and $z \in [0, 1]$

$$\mathbb{P}(f(X,S) \geq z) = \mathbb{P}(f(X,S) \geq z | S = s). \tag{2}$$

We follow Agarwal *et al.* [40] and discretise this requirement by binning the model outputs into evenly spaced bins $\mathcal{Z}$. We then require for the binned model output $f_{\text{bin}}(X, S)$ and for all $s \in S$ and $\bar{z} \in \mathcal{Z}$

$$\mathbb{P}(f_{\text{bin}}(X,s) \geq \bar{z}) = \mathbb{P}(f_{\text{bin}}(X,S) \geq \bar{z} | S = s). \tag{3}$$

In order to get a clearer insight into which regions of the output space contain disparities, we reformulate the requirement to check the probability of the intervals corresponding to each individual bin. We refer to this metric as the relative difference per bin. In our tests, we use four bins. For statistical parity the distribution of the prediction should be independent of the protected group. That is why we also check whether the shift in mean value is below a threshold as an indicator of fairness.

If a dataset has ground truth labels for emotion, we test the correctness for each group member (similar to the bounded group loss in Agarwal *et al.* [40]) and require that the difference in overall performance is low in terms of concordance correlation coefficient (CCC) and MAE for regression and in terms of unweighted average recall (UAR) and unweighted average precision (UAP) for classification.

The criterion of *Equalised Odds* [37] states that all members of a group should have equal rates for true positives and false positives in a binary classification scenario. We apply this to multi-class classification by comparing the difference in recall per class (RPC) and precision per class (PPC). For regression, we again map the model outputs into four evenly spaced bins and compare the difference in recall per bin and precision per bin. We also consider the MDE for regression in order to see if there is a bias in the mean towards a certain direction for a protected group.

For all fairness tests, we use simulations based on random models as reference for setting test thresholds. A model that outputs purely random predictions has no bias towards certain groups. However, when the number of samples in the test set is small, the change in prediction for a protected group has a higher chance of being high, even for a random model. Thus, we simulate potential test outcomes for random models under different conditions. For regression tasks, we generate random model samples from a truncated Gaussian distribution with values between 0 and 1, a mean value of 0.5, and a standard deviation of $\frac{1}{6}$. For categories, our random model samples from a uniform categorical distribution with four categories. For test metrics where the ground truth values are taken into consideration, we also generate the ground truth values randomly by sampling values from a truncated Gaussian distribution for continuous values. For categories, we consider both a uniformly distributed ground truth, as well

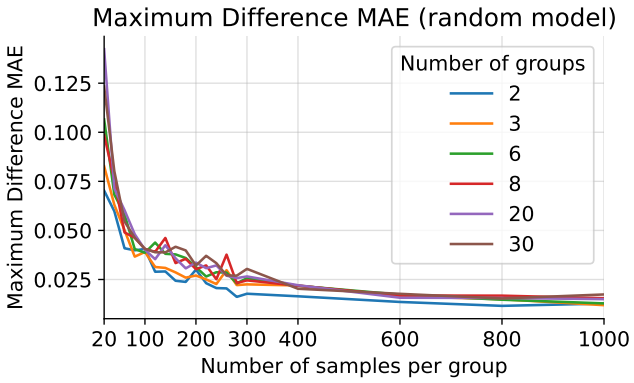

Fig. 1. Maximum difference in mean absolute error among 1000 simulations with a varying number of groups and number of samples per group.

as a sparsely distributed ground truth (with class probabilities $(0.05, 0.05, 0.3, 0.6)$). Each simulation is repeated 1000 times and we use the maximum difference in prediction for a protected group as a reference value for our test threshold. The simulation results are shown for the example of the difference in mean absolute error as test metric in Fig. 1. For certain test sets, we encountered the issue that the distribution of the ground truth for certain groups varies considerably from the distribution of other groups. The maximum difference in prediction in the simulation increases when the ground truth labels show a bias for a particular group. In order to avoid this, we balance the test sets by selecting 1000 samples from the group with the fewest samples, and 1000 samples from each other group with similar truth values. For regression tasks, this may result in certain bins having very few samples. In these cases, we decided to skip bins with too few samples. Specifically, we set the minimum number of samples n_{bin} to the expected number of samples in the first bin for a Gaussian distribution with a mean of 0.5 and a standard deviation of $\frac{1}{6}$:

$$n_{\text{bin}} = \mathbb{P}(X \leq 0.25) \cdot n, \tag{4}$$

where n is the total number of samples, and the random variable X follows the aforementioned distribution. We take the same approach for the tests with unlabelled test sets in the case that a model has very few predictions in a certain bin for the combined test set. With this, we avoid that a change in prediction of only one sample in a certain bin for one group could result in a large difference in the test metric.

Table III gives an overview of the discussed fairness tests, and also lists the applied test sets and passing conditions. All fairness thresholds are based on the previously described simulations with random models and depend on the number of protected groups in the used test, the number of samples per group, as well as whether the test data is distributed sparsely in case of categories. For some test sets certain regression bins may have fewer samples than expected for our assumption of a Gaussian distribution with a mean of 0.5 and a standard deviation of $\frac{1}{6}$ of the ground truth in labelled sets, and of the prediction in unlabelled sets. Therefore, we also list the minimum number of samples per bin n_{bin} (see Eq. 4) in parentheses behind the respective metrics. When possible, test sets with a large number of samples per protected group should be used, as the tests can lead to more meaningful conclusions. For instance, for the Fairness Accent tests, the thresholds are based on simulations with a random uniform categorical and random Gaussian model for 30 fairness groups and only 60 samples per group, and are thus relatively high. For all other tests, we can base our thresholds on simulations with at least 1000 samples per group. Note that for the Fairness Language tests, we increased the calculated thresholds slightly to accommodate for potential variations of the content and recording context across different languages in the database. The number of fairness tests are 557 for each arousal, dominance, valence, and 520 for categorical emotions. In the following paragraphs, we provide additional information to the single tests, besides what is shown in Table III.

The **Fairness Accent** tests use randomly selected samples from the speech accent archive [41], which contains more than 2000 speech samples from speakers of different nationalities and native languages. All speakers read the same paragraph in English, lasting a little over 3 minutes per recording for most cases. We use the recordings from 5 female and 5 male speakers for each accent (including native English). We compare the predictions of each of 31 accents to the predictions of the combined data.

In the **Fairness Language** tests we use 2000 randomly selected samples from Mozilla Common Voice [42] for each of the languages German, English, Spanish, French, Italian, and Mandarin Chinese. The prediction of each individual language is compared to the combined data.

It has been shown that pre-trained transformer models can use linguistic information to improve their predictions [16]. The **Fairness Linguistic Sentiment** tests investigate the effect of linguistic content for different types of languages. If the textual content does have an influence on the model predictions, it should have the same influence for each language on a fair model. To this end, we follow Triantafyllopoulos *et al.* [16] and employ CHECKLIST [43], a toolkit for generating automatic tests for natural language processing (NLP) models, including sentiment models. We expand on this by not only synthesising the English sentiment testing suite, but also generating translated versions of the test sentences for the languages German, English, Spanish, French, Italian, Japanese, Portuguese, and Mandarin Chinese. We use an OPUS-MT [44] model[1] for translation of Mandarin Chinese and ARGOS TRANSLATE [45] for all other languages, followed by manual editing to correct obvious translation errors. For the synthesis of each language, a publicly available text-to-speech model using both the libraries COQUI TTS [46] and ESPNET [47], version 0.10.6, generated the audio samples corresponding to the text. We investigate the tests for negative, neutral, and positive words in context. Up to 2000 random samples are selected for each test and each language. The prediction of the combined data is then compared to the prediction for each individual language. In this test, we only measure the influence of text sentiment for different languages

[1]https://huggingface.co/Helsinki-NLP/opus-mt-en-zh, accessed 2023/12/04

TABLE III
OVERVIEW OF THE FAIRNESS TESTS, THEIR TEST SETS, METRICS, AND PASSING CONDITIONS. FOR METRICS INVOLVING BINS, THE MINIMUM NUMBER OF SAMPLES PER BIN n_{bin} IS SHOWN IN PARENTHESES. BINS WITH FEWER THAN n_{bin} ARE SKIPPED IN THE TEST.

Test	Task	Test Sets	Metric	Condition
Fairness Accent	categories	speech accent archive	Relative Diff. Per Class	$\lvert\cdot\rvert < 0.225$
	dimensions	speech accent archive	Diff. Mean Value	$\lvert\cdot\rvert < 0.075$
			Relative Diff. Per Bin ($n_{\text{bin}} = 4$)	$\lvert\cdot\rvert < 0.225$
Fairness Language	categories	Mozilla Common Voice	Relative Diff. Per Class	$\lvert\cdot\rvert < 0.1$
	dimensions	Mozilla Common Voice	Diff. Mean Value	$\lvert\cdot\rvert < 0.03$
			Relative Diff. Per Bin ($n_{\text{bin}} = 67$)	$\lvert\cdot\rvert < 0.1$
Fairness Linguistic Sentiment	categories	CHECKLIST (synthesized)	Diff. Class Proportion Shift	$\lvert\cdot\rvert < 0.075$
	dimensions	CHECKLIST (synthesized)	Diff. Bin Proportion Shift ($n_{\text{bin}} = 67$)	$\lvert\cdot\rvert < 0.075$
			Diff. Mean Shift	$\lvert\cdot\rvert < 0.025$
Fairness Pitch	categories	MSP-Podcast (test-1)	Diff. PPC	$\lvert\cdot\rvert < 0.1$
			Diff. RPC	$\lvert\cdot\rvert < 0.225$
			Diff. UAP	$\lvert\cdot\rvert < 0.05$
			Diff. UAR	$\lvert\cdot\rvert < 0.075$
	dimensions	MSP-Podcast (test-1)	Diff. CCC	$\lvert\cdot\rvert < 0.1$
			Diff. MAE	$\lvert\cdot\rvert < 0.02$
			Diff. MDE	$\lvert\cdot\rvert < 0.04$
			Diff. Precision Per Bin ($n_{\text{bin}} = 67$)	$\lvert\cdot\rvert < 0.125$
			Diff. Recall Per Bin ($n_{\text{bin}} = 67$)	$\lvert\cdot\rvert < 0.125$
Fairness Sex	categories	IEMOCAP, MSP-Podcast (test-1)	Diff. PPC	$\lvert\cdot\rvert < 0.075$
			Diff. RPC	$\lvert\cdot\rvert < 0.175$
			Diff. UAP	$\lvert\cdot\rvert < 0.05$
			Diff. UAR	$\lvert\cdot\rvert < 0.075$
	dimensions	IEMOCAP (full), MSP-Podcast (test-1)	Diff. CCC	$\lvert\cdot\rvert < 0.075$
			Diff. MAE	$\lvert\cdot\rvert < 0.02$
			Diff. MDE	$\lvert\cdot\rvert < 0.04$
			Diff. Precision Per Bin ($n_{\text{bin}} = 67$)	$\lvert\cdot\rvert < 0.1$
			Diff. Recall Per Bin ($n_{\text{bin}} = 67$)	$\lvert\cdot\rvert < 0.1$

rather than general language biases, which are addressed in the Fairness Language tests. Therefore, we compare the shift in prediction when filtering the samples for a specific sentiment. We denote all samples with sentiment s and language l as $X_{l,s}$, and all combined samples of language l as X_l. We compute the difference between the shift in prediction for a certain sentiment and language and the average of the shifts in prediction for that sentiment for all languages $l_i, 1 \leq i \leq L$

$$\text{shift}(X_{l,s}) - \frac{1}{L}\sum_{i=1}^{L}\text{shift}(X_{l_i,s}). \tag{5}$$

By subtracting the average shift in prediction for a certain sentiment, we allow for both models that are not affected by sentiment at all and models that are affected by sentiment equally across all languages to pass the tests. For categorical emotion prediction, we compute the shift in class proportion for negative, neutral, and positive sentiment. For each class label c, a fair model's behaviour in terms of class proportion shift for one sentiment in one language should be similar to the average behaviour observed across all languages. This is tested by inserting the function shift_c, which is defined as

$$\begin{aligned}\text{shift}_c(X_{l,s}) =& \frac{1}{|X_{l,s}|}|\{y \mid y = c \text{ and } y \in \text{prediction}(X_{l,s})\}| \\ &- \frac{1}{|X_l|}|\{y \mid y = c \text{ and } y \in \text{prediction}(X_l)\}|,\end{aligned}$$

in Eq. 5. We apply the same function for dimensional emotion values, by binning the model outputs and treating the four bins as classes c. We then compute the difference in bin proportion shift analogously to the difference in class proportion shift. Additionally, we consider the shift in terms of mean value for negative, neutral, and positive sentiment. Specifically, we insert the following function $\text{shift}_{\text{mean}}$ in Eq. 5:

$$\begin{aligned}\text{shift}_{\text{mean}}(X_{l,s}) =& \text{mean}(\text{prediction}(X_{l,s})) \\ &- \text{mean}(\text{prediction}(X_l)).\end{aligned}$$

We thus compute the difference between the shift in mean value for one language and the average shift in mean value across all languages.

The **Fairness Pitch** tests address the different levels of average pitch a speaker can have. Pitch is known to be correlated with emotion, for instance, it has been observed that a higher pitch leads to higher arousal [15]. An SER model might use this correlation as a shortcut in its deductions, leading to a disparate treatment of speakers in certain pitch ranges. Consequently, we check the model behaviour for speakers of different average pitch groups on data with ground truth emotion labels and exclude speakers with fewer than 25 samples. For both categories and dimensions, we use the MSP-Podcast (MSP-Podcast) test set 1. We extract F0 frame-wise with PRAAT [48] and calculate a mean value for each segment, ignoring frames with a pitch value of 0 Hz. We exclude segments from the analysis that show an F0 below 50 Hz or above 350 Hz to avoid pitch estimation outliers to influence the tests. We then compute the average of all samples belonging to a speaker, and assign one of 3 pitch groups to that speaker. The low pitch group is assigned to speakers with an average pitch less than or equal to 145 Hz, the medium

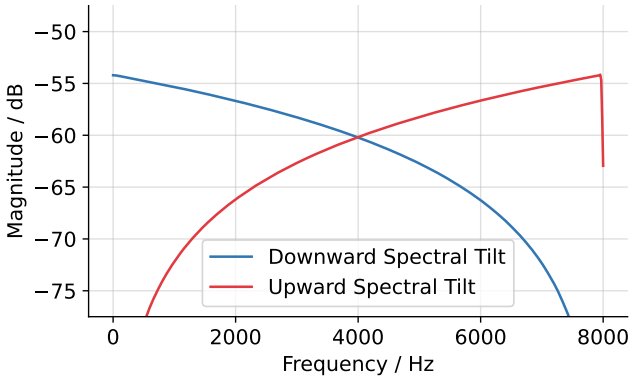

Fig. 2. Magnitude spectrum of the spectral boost downward or upward spectral tilt.

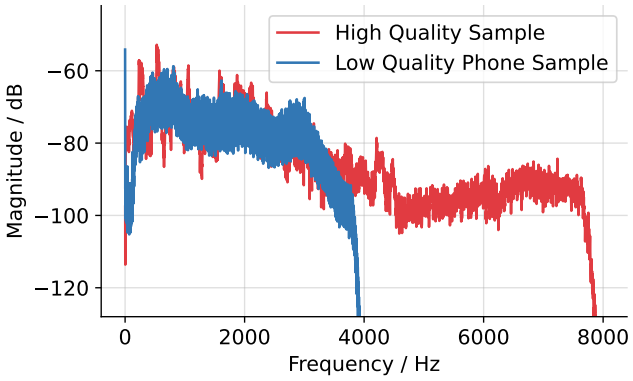

Fig. 3. Magnitude spectrum for the original and low quality phone version of an audio sample.

pitch group to speakers with an average pitch of more than 145 Hz but less than or equal to 190 Hz, and the high pitch group to speakers with an average pitch higher than 190 Hz. We compute the performance of each pitch group and compare it to the performance of the combined dataset.

For the **Fairness Sex** tests, we select test sets that have been labelled for the emotion task as well as for sex and compute the difference to the performance of the combined dataset.

D. Robustness Tests

A robust machine learning model is resilient when facing perturbations in the input data. Robustness can be evaluated by analysing how much the model predictions are affected by changes such as noise. The subcategory of adversarial robustness specifically deals with perturbations that are designed to be hard to detect and change the model's prediction (adversarial examples) [49]. We focus on applying perturbations that are likely to occur for non-malicious application scenarios rather than adversarial attacks.

When ground truth labels are available, one way to evaluate robustness is to check the difference in correctness with and without added noise [7]. For regression, we check the difference in CCC and for classification the difference in UAR and UAP. Another evaluation metric is to consider how often a perturbation changes the output, which is presented as *adversarial frequency* in Bastani *et al.* [50]. We base our metric *percentage of unchanged predictions* on this concept. For classification, the percentage of unchanged predictions is simply the percentage of samples where the class label prediction does not change from the clean audio to the audio with perturbation. It is not as straightforward to define what counts as an unchanged prediction for continuous values. For our tests, we set a threshold of 0.05, i. e., two predictions are considered to be unchanged if their absolute difference is below 0.05. This metric can be used for labelled as well as for unlabelled datasets. For regression, we additionally consider the change in average value between the clean audio and the audio with perturbations.

Table IV gives an overview of the discussed robustness tests, their test sets, and suggested passing conditions if a binary test result is desired. As with the correctness tests, the thresholds for the passing conditions are application dependent, and the thresholds in the table are only an example to show how the test results can be summarized into a percentage of passed tests. The number of robustness tests are 82 each for arousal, dominance, and valence, and 193 for categorical emotions. In the following paragraphs, we provide additional information to the single tests, besides what is shown in Table IV.

SER models have been shown to be affected by background noises although the human perception of the emotion remains the same [15]. The **Robustness Background Noise** tests investigate the robustness for various types of background noises. We simulate babble noise by mixing 4-7 speech samples from Music, Speech, and Noise Corpus (MUSAN) [51] and adding them with a signal-to-noise ratio (SNR) of 20 dB. MUSAN also contains technical and ambient sounds, as well as a music, from which one sample is added with an SNR of 20 dB for simulating environmental noise and music, respectively. We also check the effect of human coughing and sneezing sounds with samples collected by Amiriparian *et al.* [52] by adding a single cough or sneeze at a random position with an SNR of 10 dB. For artificial noise, we add white noise with an SNR of 20 dB.

The **Robustness Low Quality Phone** tests specifically target applications with audio from a low quality telephone connection. These types of recordings usually display a stronger compression, coding artifacts, and may show low pass behaviour. We mimic this by applying a dynamic range compressor, a lossy (narrow band) adaptive multi-rate (AMR) codec, and high pass filtered pink noise. Figure 3 shows the influence of the changes on the magnitude spectrum for a given audio sample.

Databases such as the Singapore English National Speech Corpus (NSC) [53] contain multiple samples of the same audio recorded simultaneously with different recording devices. We use this data in the **Robustness Recording Condition** tests and compare the predictions of the baseline recording device to the predictions of audio from alternative devices. In the case of the NSC dataset, we randomly select 5000 samples from the headset recordings and compare them to their respective

TABLE IV
OVERVIEW OF THE ROBUSTNESS TESTS, THEIR TEST SETS, METRICS, AND PASSING CONDITIONS.

Test	Task	Test Sets	Metric	Condition
Robustness Background Noise	categories	CREMA-D, EMOVO, IEMOCAP, MELD, MSP-Podcast (test-1)	Change UAP	> -0.05
			Change UAR	> -0.05
			Perc. Unchanged Predictions	> 0.9
	dimensions	IEMOCAP, MSP-Podcast (test-1)	Change Average Value	$\lvert\cdot\rvert < 0.03$
			Change CCC	> -0.05
			Perc. Unchanged Predictions	> 0.9
Robustness Low Quality Phone	categories	CREMA-D, EMOVO, IEMOCAP, MELD, MSP-Podcast (test-1)	Change UAP	> -0.05
			Change UAR	> -0.05
			Perc. Unchanged Predictions	> 0.5
	dimensions	IEMOCAP, MSP-Podcast (test-1)	Change Average Value	$\lvert\cdot\rvert < 0.05$
			Change CCC	> -0.05
			Perc. Unchanged Predictions	> 0.5
Robustness Rec. Condition	categories, dimensions	NSC	Perc. Unchanged Predictions	> 0.8
Robustness Sim. Rec. Condition	categories, dimensions	EMOVO, NSC, TIMIT	Perc. Unchanged Predictions	> 0.8
Robustness Small Changes	categories	CREMA-D, EMOVO, IEMOCAP, MELD, MSP-Podcast (test-1)	Perc. Unchanged Predictions	> 0.95
	dimensions	IEMOCAP, MSP-Podcast (test-1)	Perc. Unchanged Predictions	> 0.95
Robustness Spectral Tilt	categories	CREMA-D, EMOVO, IEMOCAP, MELD, MSP-Podcast (test-1)	Change UAP	> -0.02
			Change UAR	> -0.02
			Perc. Unchanged Predictions	> 0.8
	dimensions	IEMOCAP, MSP-Podcast (test-1)	Change Average Value	$\lvert\cdot\rvert < 0.03$
			Change CCC	> -0.05
			Perc. Unchanged Predictions	> 0.8

recordings using the boundary microphone, as well as to their respective mobile phone recordings, and compute the percentage of unchanged predictions.

Another option to evaluate robustness for different recording conditions is to simulate them. The **Robustness Simulated Recording Condition** tests simulate audio recordings at different locations using impulse responses from the Multichannel Acoustic Reverberation Database at York (MARDY) [54] dataset, and different rooms using impulse responses from the Aachen Impulse Response (AIR) [55] dataset. We use the impulse response in the centre position at 1 meter distance as the baseline to test robustness to other positions. For the room test, we use the impulse response of a recording booth as reference and compare to impulse responses of other rooms recorded at similar distances as the reference. We apply the impulse responses to Italian Emotional Speech Database (EMOVO), to 5000 randomly selected headset recordings from the NSC dataset, and to 5000 randomly selected samples from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) [56], which contains broadband recordings of 630 speakers reading ten sentences in American English. We select those three datasets as they provide dry speech recordings with a high SNR. Then, we measure the percentage of unchanged predictions.

The **Robustness Small Changes** tests apply very small transformations to the audio that were designed to be perceived as subtle to a person. Each of the small augmentations described in the following is compared to the baseline. The Additive Tone test adds a sinusoid with a randomly selected frequency between $5000\,\mathrm{Hz}$ and $7000\,\mathrm{Hz}$ with an SNR of $40\,\mathrm{dB}$, $45\,\mathrm{db}$, or $50\,\mathrm{dB}$. The Append Zeros and Prepend Zeros tests add 100, 500, or 1000 samples containing zeros to the end of the signal, or respectively the beginning of the signal. The Clip test clips 0.1%, 0.2%, or 0.3% of the input sample. The Crop Beginning and Crop End tests remove 100, 500, or 1000 samples from the beginning or end of the signal respectively. The Gain test changes the gain of the signal by a value randomly selected from $-2\,\mathrm{dB}$, $-1\,\mathrm{dB}$, $1\,\mathrm{dB}$, and $2\,\mathrm{dB}$. The Highpass Filter and Lowpass Filter tests apply a Butterworth filter of order 1 with a cutoff frequency of $50\,\mathrm{Hz}$, $100\,\mathrm{Hz}$, or $150\,\mathrm{Hz}$ for a lowpass, or $7500\,\mathrm{Hz}$, $7000\,\mathrm{Hz}$, or $6500\,\mathrm{Hz}$ for a highpass. The White Noise test adds Gaussian distributed white noise with a root mean square based SNR randomly selected from $35\,\mathrm{dB}$, $40\,\mathrm{dB}$, and $45\,\mathrm{dB}$.

The **Robustness Spectral Tilt** tests simulate the boosting of low or high frequencies in the spectrum. We simulate such spectral tilts by attenuating or emphasising the signal linearly, while ensuring that the overall signal level stays the same if possible without clipping. Figure 2 shows the magnitude spectrum of the applied upward and downward spectral tilt.

III. MODELS

We test all models trained by Wagner *et al.* [57], namely a 14-layer Convolutional Neural Network (*CNN14*), which was not pre-trained; four models based on the HuBERT [58] and wav2vec 2.0 [59] architectures, each pre-trained on English audiobooks, using 12 transformer layers (*hubert-b* and *w2v2-b*), or 24 transformer layers (*hubert-L* and *w2v2-L*), where hubert-L has been pre-trained on the Libri-Light corpus [60] and the other three have been pre-trained on the LibriSpeech corpus [61]; *w2v2-L-robust*, a model identical to *w2v2-L*, but instead pre-trained on English recordings of audiobooks (Libri-Light [60]), Wikipedia sentences (Common Voice [42]), and telephone speech (Switchboard [62] and Fisher [63]); *w2v2-L-vox*, a model identical to *w2v2-L* but instead pre-trained on parliamentary speech in multiple languages (Vox-

TABLE V
PERCENTAGE OF PASSED TESTS FOR EACH OF THE EMOTION PREDICTION TASKS (AROUSAL, DOMINANCE, VALENCE, EMOTIONIAL CATEGORIES AND AS AVERAGE ∅ OVER THOSE FOUR.

Model	A	D	V	C	∅
All tests					
w2v2-L-robust	.871	**.854**	.805	.771	**.825**
hubert-L	.859	.847	.794	**.785**	.821
w2v2-L-vox	**.875**	.844	.811	.750	.820
hubert-b	.866	.853	**.812**	.724	.814
w2v2-L-xls-r	.854	.842	.776	.764	.809
w2v2-L	.870	.817	.794	.745	.806
w2v2-b	.837	.839	.778	.743	.799
CNN14	.796	.815	.739	.727	.769
Correctness tests					
hubert-L	**.890**	.730	.583	**.571**	**.694**
w2v2-L-robust	.841	.769	**.596**	.554	.690
w2v2-L	.880	**.787**	.436	.488	.648
w2v2-L-xls-r	.835	.782	.411	.382	.602
hubert-b	.741	.701	.492	.452	.596
w2v2-b	.689	.719	.377	.493	.570
w2v2-L-vox	.758	.752	.361	.386	.564
CNN14	.625	.515	.281	.390	.453
Fairness tests					
w2v2-L-vox	.947	.945	.982	.948	**.955**
hubert-b	.958	**.970**	.972	.911	.953
w2v2-L-xls-r	.944	.939	.943	.976	.950
w2v2-L	**.964**	.900	**.994**	.917	.944
hubert-L	.917	.953	.944	.941	.939
w2v2-L-robust	.915	.923	.930	.952	.930
w2v2-b	.921	.921	.927	.948	.929
CNN14	.942	.933	.844	**.988**	.927
Robustness tests					
w2v2-L-robust	**.714**	**.788**	.567	.750	**.705**
hubert-L	.613	.668	**.598**	.796	.669
w2v2-L-vox	.705	.580	.430	**.822**	.634
hubert-b	.434	.582	.395	.593	.501
w2v2-L-xls-r	.400	.534	.364	.704	.500
w2v2-L	.442	.483	.462	.562	.487
w2v2-b	.516	.469	.310	.480	.444
CNN14	.259	.282	.209	.467	.304

Populi [64]); and *w2v2-L-xls-r*, a model identical to *w2v2-L* but instead pre-trained on more than 400k hours across all domains and multiple languages (VoxPopuli [64], ML LibriSpeech [65], Common Voice [42], VoxLingua107 [66], and BABEL [67]). All of them are fine-tuned on the MSP-Podcast corpus (v1.7) [29] with the multitask of the three dimensions of arousal, dominance, and valence. For each of the listed dimensional models, we also train a categorical counterpart with the four classes anger, happiness, neutral, and sadness. We start with the same pre-trained models and fine-tune them on the same MSP-Podcast data, but use the categorical labels as targets instead.

IV. RESULTS

None of the models passes all tests as indicated by the percentage of passed tests in Table V for the four different tasks arousal, dominance, valence, and emotional categories. The percentage of passed test results are computed as the average percentage over the involved tests, to better reflect that single tests differ in the number of involved metrics and test sets.

The test results allow a comparison of the different models. The models can also be ranked based on the percentage of passed tests, but this has to be done carefully as different tests might have different importance for particular applications. Further, the single tests, or tests grouped by correctness, fairness, robustness can provide insights into model behaviour.

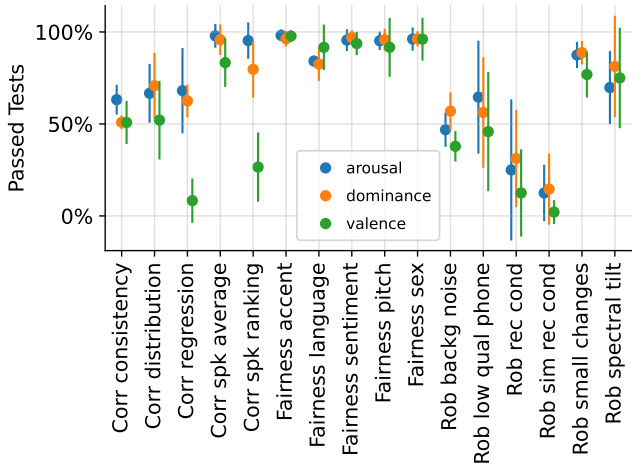

Fig. 4. Percentage of passed tests averaged over all models presented with standard deviation for all tests involving a dimensional emotion task. Corr stands for Correctness, Rob for Robustness, spk for speaker, backg for background, qual for quality, rec for recording, and cond for condition.

w2v2-L-robust passes the most tests for all tasks. All wav2vec2 and HuBERT based models are ranked before the *CNN14* baseline and larger model architectures tend to pass more tests.

Detailed results with additional plots are available under https://audeering.github.io/ser-tests/. In the following subsections, we focus on a few interesting results.

A. Correctness

The test results for correctness indicate that the valence task is much harder for the models, compare Figure 4. For arousal and dominance, a model passes on average 68% and 63% of the Correctness Regression tests, for valence the average is only 8%.

The Correctness Consistency tests estimate how well the models' arousal, dominance, and valence predictions fit for a sample with an assigned emotional category as ground truth. *w2v2-L*, *w2v2-L-vox*, and *w2v2-L-xls-r* are the most consistent. *hubert-L* and *w2v2-L-robust* pass a similar number of tests for arousal and dominance, but are less consistent for valence, where they tend to predict the same valence value independent of the underlying categorical emotion. Results for *hubert-L* are shown in Figure 5.

B. Fairness

All models pass at least 91% of the fairness tests for any task. The models passing the most tests are *w2v2-L-vox*, *hubert-b*, and *w2v2-L-xls-r*, on average passing around 95% of the tests. There is no single fairness test that is completely failed by a model. For most of the models, the lowest percentage of passed tests occurs for the Fairness

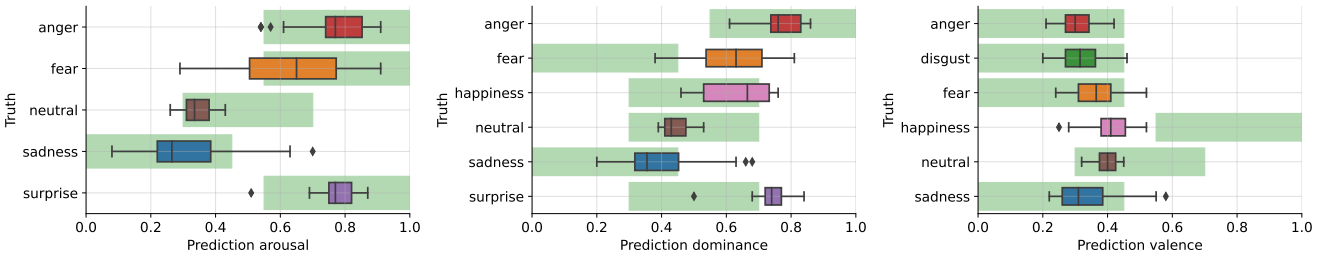

Fig. 5. Predictions of *hubert-L* for arousal (left), dominance (centre), valence (right) on the RAVDESS test set, split by the categorical emotions the samples are annotated for in RAVDESS. The green area marks the region in which a dimensional prediction would be rated as consistent with the annotated emotional category by the Correctness Consistency tests.

Language tests. There, all wav2vec2 based models show a tendency to return lower values for English and higher values for Italian for all three emotional dimensions. The HuBERT based models show a tendency for higher values for Italian on all three emotional dimensions.

w2v2-L-vox is the model least influenced by linguistic sentiment. It is the only model passing all Fairness Linguistic Sentiment tests for arousal, dominance, and valence. *hubert-b*, *w2v2-b*, *hubert-L*, and *w2v2-L-robust* clearly show a strong influence by sentiment when predicting valence and categorical emotion for English. A positive sentiment leads to higher valence values, a higher number of *happiness* and a reduced number of *sadness* predictions, and vice versa. Interestingly, the same trend is observable with *w2v2-L-robust* and *hubert-L* for arousal and dominance as well.

C. Robustness

The robustness tests roughly group the models in three classes. *hubert-L* and *w2v2-L-robust* are more robust than all other models, *CNN14* is less robust than all other models, and the remaining models pass a similar number of robustness tests.

There is a difference between *hubert-L* and *w2v2-L-robust*, as the latter is more robust for arousal and dominance whereas the former is more robust for valence and emotional categories. In both cases, there is no single robustness test that explains the difference, but nearly all tests contribute to it.

Most of the models fail more than 50% of all Robustness Background Noise tests, and are mainly affected by coughing, sneezing, and white noise. When adding sneezing or coughing sounds to the input signal when predicting emotional categories, the predictions of the models shifts towards the *happiness* class, see Figure 6 for results for *hubert-L*, indicating that coughing and sneezing might be confused with laughter by the models.

V. Discussion

Machine learning models for the speech emotion recognition tasks are usually benchmarked based on their performance in terms of CCC for dimensional tasks or UAR for categorical tasks. When calculating the average over CCC and UAR of the models over all four tasks, *w2v2-L-robust* and *hubert-L* are the best models mainly due to their performance for valence, *CNN14* is the worst model, and all other models are very similar, compare Table VI. The results of the tests show accordingly the most passed tests for *w2v2-L-robust* and *hubert-L*, indicating that for the tested models the standard benchmarks are indeed able to select the overall best models.

The advantage of tests in addition to benchmarks is that they are better at characterising model behaviour, and allow the exclusion of a model from application until it passes certain tests that might be critical for the application. Even though *w2v2-L-robust* and *hubert-L* are the best models regarding correctness and robustness, they fail several robustness tests that add background noise or use different recording conditions. Hence, they might not be suited for real world applications without further augmentations during pre-training, fine-tuning, or knowledge distillation [68]. *w2v2-L-robust* and *hubert-L* also fail fairness tests, which indicates the models might have small biases regarding sex and language.

Triantafyllopoulos *et al.* [16] showed that some of the

TABLE VI
Average CCC for arousal (**A**), dominance (**D**), and valence (**V**), and average UAR for emotional categories (**C**). The last column presents the average (∅) over the four values and the models are ranked based on this average.

Model	A	D	V	C	∅
w2v2-L-robust	**.64**	**.54**	**.50**	**.58**	**.57**
hubert-L	.63	.53	.49	.55	.55
w2v2-L-vox	.63	**.54**	.36	.50	.51
hubert-b	.60	.52	.40	.51	.51
w2v2-b	.60	.53	.37	.52	.51
w2v2-L	.63	.53	.32	.53	.50
w2v2-L-xls-r	.63	.53	.30	.50	.49
CNN14	.50	.39	.18	.44	.38

TABLE VII
Change of the predicted average valence value for neutrally spoken words in English with **negative**, **neutral**, and **positive** sentiment.

Model	negative	neutral	positive
hubert-L	−.12	+.02	+.12
w2v2-L-robust	−.12	±.00	+.13
hubert-b	−.09	−.03	+.11
w2v2-b	−.03	−.02	+.05
w2v2-L-vox	−.02	±.00	+.02
w2v2-L	±.00	+.01	±.00
w2v2-L-xls-r	±.00	+.01	±.00
CNN14	−.01	+.02	±.00

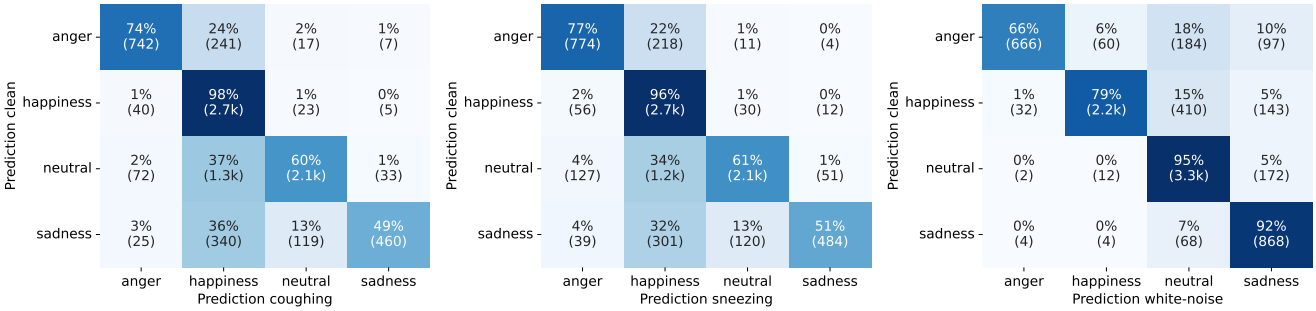

Fig. 6. Confusion matrices for the prediction of emotional categories by *hubert-L* on the clean MSP-Podcast test set 1 comparing to the MSP-Podcast test set 1 when adding coughing with an SNR of 10 dB (left), sneezing with an SNR of 10 dB (centre), or white noise with an SNR of 20 dB (right).

success in predicting valence can be attributed to the linguistic knowledge encoded in the self-attention layers of the wav2vec2 or HuBERT models. The results for the correctness tests for valence (Table V) and the CCC values for valence (Table VI) indicate that *w2v2-L-robust*, *hubert-L*, and *hubert-b* are showing the best performance and hence might rely more on linguistic content. This hypothesis can be checked by evaluating how valence is influenced by the sentiment of the spoken text. This is measured as part of the Fairness Linguistic Sentiment test, which synthesises neutrally spoken words/sentences with negative, neutral, or positive sentiment. Table VII lists the shift in average predicted valence for the different sentiment conditions. *hubert-L* and *w2v2-L-robust*, followed by *hubert-b*, are the models that show the largest shift towards lower valence for negative sentiment and towards higher valence for positive sentiment. Whereas the models *w2v2-L*, *w2v2-L-xls-r*, and *w2v2-L-vox* show no, or only a small shift. The results indicate that the best performing models for valence do indeed take sentiment into account.

As *hubert-b* and *w2w2-b* and *hubert-L* and *w2v2-L* are trained on the same data, respectively, this indicates that HuBERT might be better suited to learn the linguistic information. On the other hand, *w2v2-L-robust* was also able to learn linguistic information based on a similar amount of data as *hubert-L*. What seems to reduce the ability to learn lingustic information is the inclusion of different languages in the training data as is the case for *w2v2-L-xls-r* which was trained on ~ 7 times the amount of data than *w2v2-L-robust*, but includes several different languages.

VI. Conclusion

We proposed a large set of 3,062 different tests to judge the behaviour of speech emotion recognition models in terms of correctness, fairness, and robustness on the tasks of predicting arousal, dominance, valence, and categorical emotions like anger. The tests allow to request a certain amount of correctness or robustness depending on the desired application of the models. We further provide an approach to estimate test thresholds automatically for testing model fairness.

When applying the test suite to a selection of eight models all trained on MSP-Podcast, the results show that the number of overall passed tests of the models relates to rankings on classical speech emotion recognition benchmarks based on CCC and UAR, indicating that a more correct model seems also to perform better or as well in fairness and robustness related tests. Still, the results also show that all models have slight biases for sex and language, and that they are not robust enough for applications that involve different microphones or background noise. Hence, the tests directly indicate how to further improve those models.

Furthermore, detailed test results provide insights into how the models are able to reach their given performance, or what input factors influence model results. For example, the results of the Fairness Linguistic Sentiment tests show that the three best performing models for valence are all achieving this by relying on sentiment of the spoken English text.

VII. Acknowledgements

The authors would like to thank A. Triantafyllopoulos, C. Oates, and A. Hvelplund for their valuable discussions and feedback during the development of the tests, and J. Wagner for training some of the models tested in this paper.

VIII. References

[1] J. Thiyagalingam, M. Shankar, G. Fox, and T. Hey, "Scientific machine learning benchmarks," *Nature Reviews Physics*, vol. 4, no. 6, pp. 413–420, 2022.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[3] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446.

[4] J. P. Turian *et al.*, "HEAR 2021: Holistic evaluation of audio representations," *arXiv preprint arXiv:2203.03022*, 2022.

[5] A. D'Amour *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *Journal of Machine Learning Research*, vol. 23, pp. 1–61, 2022.

[6] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: Association for Computing Machinery, 2019, pp. 220–229. DOI: 10.1145/3287560.3287596.

[7] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Transactions on Software Engineering*, vol. 48, no. 1, pp. 1–36, 2020. DOI: 10.1109/TSE.2019.2962027.

[8] P. Ammann and J. Offutt, *Introduction to software testing*. Cambridge University Press, 2016.

[9] C. Murphy, G. E. Kaiser, and M. Arias, "An approach to software testing of machine learning applications," in *Proceedings of the Nineteenth International Conference on Software Engineering & Knowledge Engineering (SEKE)*, Boston, MA, USA: Knowledge Systems Institute Graduate School, 2007, pp. 167–172.

[10] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the 40th international conference on software engineering*, 2018, pp. 303–314.

[11] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," in *Association for Computational Linguistics (ACL)*, 2020.

[12] H. Zhou, W. Li, Z. Kong, J. Guo, Y. Zhang, B. Yu, L. Zhang, and C. Liu, "Deepbillboard: Systematic physical-world testing of autonomous driving systems," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, 2020, pp. 347–358.

[13] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, *et al.*, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech & Language*, vol. 53, pp. 156–180, 2019.

[14] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7697–7701. DOI: 10.1109/ICASSP43922.2022.9747348.

[15] M. Jaiswal and E. M. Provost, "Best practices for noise-based augmentation to improve the performance of emotion recognition "in the wild"," *arXiv preprint arXiv:2104.08806*, 2021.

[16] A. Triantafyllopoulos, J. Wagner, H. Wierstorf, M. Schmitt, U. Reichel, F. Eyben, F. Burkhardt, and B. W. Schuller, "Probing speech emotion recognition transformers for linguistic knowledge," in *Interspeech 2022, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds., 2022, pp. 146–150. DOI: 10.21437/interspeech.2022-10371.

[17] M. Schmitz, R. Ahmed, and J. Cao, "Bias and fairness on multimodal emotion detection algorithms," *arXiv preprint arXiv:2205.08383*, 2022.

[18] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[19] Z. Ren, T. T. Nguyen, Y. Chang, and B. W. Schuller, "Fast yet effective speech emotion recognition with self-distillation," *arXiv preprint arXiv:2210.14636*, 2022.

[20] H. Spieker and A. Gotlieb, "Towards testing of deep learning systems with training set reduction," *arXiv preprint arXiv:1901.04169*, 2019.

[21] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[22] H. Wierstorf and A. Derington, *Test splits for crema-d, emodb, iemocap, meld, ravdess*, 2023. DOI: 10.5281/zenodo.10229583.

[23] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database," in *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1997, pp. 1695–1698. DOI: 10.21437/Eurospeech.1997-482.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech.," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, vol. 5, Lisbon, Portugal: ISCA, 2005, pp. 1517–1520.

[25] G. Costantini, I. Iaderola, A. Paoloni, M. Todisco, *et al.*, "Emovo corpus: An italian emotional speech database," in *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, European Language Resources Association (ELRA), 2014, pp. 3501–3504.

[26] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[27] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[28] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "Emotionlines: An emotion corpus of multi-party conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[29] 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, 2019.

[30] *Polish emotional speech database*, Retrieved from http://www.eletel.p.lodz.pl/bronakowski/med_catalog/, Mar. 27, 2020.

[31] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, e0196391, 2018.

[32] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007. DOI: 10.1111/j.1467-9280.2007.02024.x.

[33] C. Gillioz, J. R. Fontaine, C. Soriano, and K. R. Scherer, "Mapping emotion terms into affective space," *Swiss Journal of Psychology*, 2016.

[34] H. Hoffmann, A. Scheck, T. Schuster, S. Walter, K. Limbrecht, H. C. Traue, and H. Kessler, "Mapping discrete emotions into the dimensional space: An empirical approach," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2012, pp. 3316–3320.

[35] G. K. Verma and U. S. Tiwary, "Affect representation and recognition in 3d continuous valence–arousal–dominance space," *Multimedia Tools and Applications*, vol. 76, pp. 2159–2183, 2017.

[36] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[37] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, 2021, ISSN: 0360-0300. DOI: 10.1145/3457607.

[38] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[39] E. Del Barrio, P. Gordaliza, and J.-M. Loubes, "Review of mathematical frameworks for fairness in machine learning," *arXiv preprint arXiv:2005.13755*, 2020.

[40] A. Agarwal, M. Dudík, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *International Conference on Machine Learning*, PMLR, 2019, pp. 120–129.

[41] S. Weinberger, *Speech accent archive*, Retrieved from http://accent.gmu.edu, 2015.

[42] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[43] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," in *Association for Computational Linguistics (ACL)*, 2020.

[44] J. Tiedemann and S. Thottingal, "OPUS-MT — Building open translation services for the World," in *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

[45] P. Finlay and C. Argos Translate, *Argos translate*, version 1.8.0, Feb. 4, 2023. [Online]. Available: https://github.com/argosopentech/argos-translate.

[46] G. Eren and T. C. T. Team, *Coqui-ai/TTS, version 0.6.1*, Jan. 2021. DOI: 10.5281/zenodo.6334862. [Online]. Available: https://www.coqui.ai.

[47] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7654–7658.

[48] P. Boersma, *Praat: Doing phonetics by computer [computer program]*, Retrieved from http://www.praat.org/, 2023.

[49] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[50] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, and A. Criminisi, "Measuring neural net robustness with constraints," *Advances in neural information processing systems*, vol. 29, 2016.

[51] D. Snyder, G. Chen, and D. Povey, *MUSAN: A Music, Speech, and Noise Corpus*, arXiv:1510.08484v1, 2015. eprint: 1510.08484.

[52] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. W. Schuller, "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 340–345. [Online]. Available: https://doi.org/10.1109/ACII.2017.8273622.

[53] J. X. Koh, A. Mislan, K. Khoo, B. Ang, W. Ang, C. Ng, and Y.-Y. Tan, "Building the Singapore English National Speech Corpus," in *Proc. Interspeech 2019*, 2019, pp. 321–325. DOI: 10.21437/Interspeech.2019-1525.

[54] J. Y. Wen, N. D. Gaubitch, E. A. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the mardy database," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006, pp. 1–4.

[55] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," English, in *Proceedings of International Conference on Digital Signal Processing (DSP)*, IEEE, IET, EURASIP, Santorini, Greece: IEEE, Jul. 2009, pp. 1–4, ISBN: 978-1-42443-298-1.

[56] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Timit acoustic-phonetic continuous speech corpus," 1983. DOI: https://doi.org/10.35111/17gk-bn40.

[57] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, 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*, pp. 1–13, 2023.

[58] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[59] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 12 449–12 460.

[60] and others, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7669–7673.

[61] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[62] J. Godfrey and E. Holliman, "Switchboard-1 release 2 ldc97s62," *Linguistic Data Consortium*, p. 34, 1993.

[63] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher english training speech part 1 transcripts," *Philadelphia: Linguistic Data Consortium*, 2004.

[64] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80.

[65] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," 2020.

[66] J. Valk and T. Alumäe, "Voxlingua107: A dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 652–658.

[67] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, International Speech Communication Association (ISCA), 2014, pp. 16–23.

[68] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015, NIPS 2014 Deep Learning Workshop.