IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 31, 2021, accepted September 8, 2021, date of publication September 13, 2021,
date of current version September 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3112165

Marginal Effects of Language and Individual Raters on Speech Quality Models

MICHAEL CHINEN, **(Member, IEEE)**
Google, San Francisco, CA 94105, USA
e-mail: mchinen@google.com

ABSTRACT Speech quality is often measured via subjective testing, or with objective estimators of mean opinion score (MOS) such as ViSQOL or POLQA. Typical MOS-estimation frameworks use signal level features but do not use language features that have been shown to have an effect on opinion scores. If there is a conditional dependence between score and language given these signal features, introducing language and rater predictors should provide a marginal improvement in predictions. The proposed method uses Bayesian models that predict the individual opinion score instead of MOS. Several models that test various combinations of predictors were used, including predictors that capture signal features, such as frequency band similarity, as well as features that are related to the listener, such as a language and rater index. The models are fit to the ITU-T P. Supplement 23 dataset, and posterior samples are drawn from distributions of both the model parameters and the resulting opinion score outcomes. These models are compared to MOS models by integrating over posterior samples per utterance. An experiment was conducted by ablating different predictors for several types of Bayesian hierarchical models (including ordered logistic and truncated normal individual score distributions, as well as MOS distributions) to find the marginal improvement of language and rater. The models that included language and/or rater obtained significantly lower errors (0.601 versus 0.684 root-mean-square error (RMSE)) and higher correlation. Additionally, individual rater models matched or exceeded the performance of MOS models.

INDEX TERMS Bayesian models, causality, culture, language, mean opinion score, speech quality estimation, subjective testing.

I. INTRODUCTION

Measuring and estimating speech quality is an important task for many fields. For example, in speech synthesis and coding [1], subjective measurements of quality can be used to validate novel designs, and may be especially useful when traditional objective metrics like SNR diverge from human perception. The absolute categorical ranking (ACR) test asks raters to measure the quality of speech utterances under various test conditions by assigning a score from 1 (bad) to 5 (excellent), with recommendations for conducting the test in ITU P. 800 [2]. Typically, each utterance has multiple listeners. The mean opinion score (MOS) can be calculated by aggregating the scores over each utterance or all the utterances within a given condition. MOS is a standard measurement that is used in research and development of many speech applications such as codecs and speech enhancement [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Masini.

Because it is expensive and logistically challenging to conduct a subjective experiment, researchers and developers often use estimates of MOS that do not require running a subjective test. Some MOS estimation techniques estimate MOS using a model of the effects of different psychologically pertinent factors (e.g. echo, delay, SNR, language, gender), without looking at the actual signal [4], [5]. Tools such as ViSQOL [6]–[8], POLQA [9], [10], PESQ [11], [12], and the E-model [13], [14] provide immediate and objective estimates of MOS using an intrusive method that looks mainly at the signal as opposed to these factors (by considering both the reference and degraded signal). These models obtain a MOS by fitting a mapping function using features from the signals and MOS extracted from datasets such as ITU-T P Supplement 23 [15] by aggregating all listeners over each utterance.

Additionally, there are deep learning methods for estimating MOS non-intrusively, which learn latent features that are mapped to MOS [16] by looking even further into the

lower level aspects of the signal. Deep learning requires significant data, and some work has been done to bootstrap by augmenting the data [17] or by clustering [18]. Another approach may be to use data more efficiently, by looking at the causes of measurement and sampling error in individual scores. The trend in research is to use machine learning to extract increasingly useful information from the signal. The vast majority of MOS estimation tools use only-signal level predictors, which effectively treats all populations of human raters as identical, which is clearly not the case.

Identifying and employing interesting information in the data is important to be able to design a good model. When the listeners are aggregated for fitting a mapping function to MOS, (e.g. by taking the arithmetic mean of all listeners), information about the variance of the scores with respect to the utterances is discarded, as well as information about the individuals. The resulting models are therefore unable to describe the effect of the presence of individual listeners. As an extreme example, if a very optimistic rater rated everything a '5', such a model is not able to capture the fact that the presence of the rater causes a slightly higher mean score and will instead have a higher error in predictions. The result is that the individuals that are outliers have an oversized effect on the mapping.

Proper handling outliers properly is increasingly important with crowd-sourced data, e.g. listening tests that are conducted over a web service such as Amazon Mechanical Turk. In these cases the quality of ratings can be worse than traditional lab tests and may require additional pre-screening and restrictions [19], (e.g. raters that always give the highest or lowest score due to any number of reasons including bad headphones or malicious ratings), and outliers will have an undesirably large influence on the mapping. The ITU provides post filtering recommendations that filter based on deviation from normal behavior, which generally works well. However, the filtering process is sensitive and can have undesirable consequences. Such filtering has the potential to remove genuine scores and leave undesirable ratings in the data.

A model of the distribution of individual rater scores that is aware of the bias that a particular rater has over multiple utterances should provide the ability to exclude undesirable raters in a more systematic fashion, as well as extracting more information in raters that have significant bias. Post filtering will exclude some raters that simply have a positive or negative bias, although their relative ratings match the overall trends. Modelling the individual listener score will allow for the model to be able to take into account this rater data, accounting for the rater bias. Additionally, many researchers have pointed out issues with using MOS as the primary quality metric and have proposed alternatives [20]–[22]. Modeling the individual rater score allows using the model for other metrics that are alternatives or complements to MOS.

There are numerous other biases at work in a subjective audio quality test, including many biases that are related to how the test is conducted [23]. Besides the audio signal, test environment, and individual listener bias, the effect of language and culture may have a causal relationship with opinion scores. Some languages and cultures rate the same set of test conditions higher or lower than others. For example, Japanese listeners tend to rate the quality of speech lower on average [24], and have less variance in their ratings than listeners with other native languages. As another example where Japanese means are lower, Figure 3 depicts Japanese and French ratings in the ITU-T P Supplement 23 dataset with the same test conditions. This effect appears to be cultural, or at least it is not exclusively attributable to phonetic differences in languages, as other quality of experience (QoE) ratings from other non-speech domains (e.g. restaurant ratings) for the Japanese seem to follow this trend [25]. To further confound this issue, the quality labels recommended in [2] (such as 'excellent'/'good'/'fair'/'poor'/'bad') may not imply a linear progression of quality within a language, and furthermore, it is not likely that each level has equivalence between two different languages [23].

All of these factors suggest that for the same set of conditions, a certain difference in scores is expected for different languages. This difference means that a model that does not have language as one of the model's predictors will produce a larger error on the predicted MOS when compared to a similar model that has the language predictor. Furthermore, for users of a MOS estimation tool, it may be more desirable to estimate the MOS for the language that is being tested, as opposed to the global MOS for all listeners from any language. Typically MOS tests require that the listeners and utterances use the same language, so it should not be expected of the model to perform well across languages unless extra consideration or data is provided. Lastly, it should be considered that language, culture, technology and perception are not frozen objects, but are constantly evolving and interacting. This hints that a model should consider more than the signal alone, such as attributes of the raters.

Opinion score data is scarce for quality assessment problems compared to other fields such as speech synthesis, because human rating data does not occur as naturally as raw speech and usually must be collected manually by conducting expensive tests. Bayesian models provide a solution to deal with limited and outlying data by using prior distributions to provide a baseline that can be updated instead of fully depending only on the observations. For example, there are often 24 listeners per utterance. If all of them happen to rate the score as a '5', it is not reasonable to assume there is zero variance and all future raters will rate it a '5' as well with total confidence. To handle these problems, we propose to use a Bayesian hierarchical model of an ordered categorical distribution to model individual opinion scores based on speech, listener, and language features.

The experiments for this paper use a Bayesian model on individual rater score instead of a MOS-based model with no loss of generality. That is, such a model can be used to compute anything that a MOS-based model can. In addition,

to provide models that are closer to real-world MOS estimators, we propose to fit models that include signal level predictors, to provide an indication of the marginal benefit of adding language and rater predictors.

The main contributions of this paper are as follows:

- We fit a Bayesian model with ordered categorical distributions and truncated normal distributions to model individual scores, and compare these to mean opinion score models.
- We measure the effects of language and listener in the observations and the posterior samples.
- We compare various models and show that a language and listener-aware model have significantly lower error than the model without it, even after adding signal features.

An explicit non-goal should be stated: this paper is not concerned with finding the best model and predictors that produce the lowest error. Instead, it uses relatively simple models that can be easily compared. The findings from this paper should be useful for model design and input feature selection in other frameworks including existing MOS estimation tools and deep learning-based models that attempt to find the most accurate model.

This paper is presented as follows: first, related work is described. The next section describes the model, including the choice of individual scores, a causal model, and the specific Bayesian models that are considered. The next section describes the experiments, dataset, and results. A concluding section summarizes the paper.

II. RELATED WORK

There has been work on using hierarchical Bayesian models of MOS that consider the heterogeneity of the data, evaluating the effect of speaker gender across multiple languages [5] and loudness patterns [26]. This work showed the effectiveness of Bayesian modelling given no signal level predictors. Signal predictors have allowed existing frameworks like ViSQOL or POLQA to predict MOS with an even higher accuracy. The next question to ask is whether there is a marginal benefit to adding language predictors.

Previous work has explored using multinomial models to model the distribution of all rater scores [20]. The multinomial distribution has advantages over modelling MOS directly and is suitable for certain applications such as using the distribution directly as a quality measure instead of MOS (and is able to generalize to a MOS-model). However, it does not capture information about the individual raters.

As mentioned in previous work [8], [20], point estimates are strictly less useful than distributions of the scores, which provide a measure of uncertainty. The uncertainty is useful for both the end user, who may wish to know the range of scores to expect, but also to conduct statistical tests or compute probabilistic metrics. This is a separate question from whether or not to use individual scores or per-utterance MOS data in the model. Bayesian models are always probabilistic, so they always have a notion of uncertainty.

III. MODEL DESIGN

In this section, the properties of a desirable model for opinion scores are discussed. The design of the model includes not only considering the type of model (e.g. Bayesian, deep learning, or linear regression), but also the options for what quantity to model (e.g. individual score or mean score), and what predictors to use (e.g. signal-level features and language metadata) given a hypothetical causal relationship between both the predictors and the outcomes.

A. INDIVIDUAL OPINION SCORE VS. MEAN OPINION SCORE

A model of individual opinion score is strictly more useful than a model of mean opinion score, because a mean opinion score can be calculated by grouping the posterior samples of the individual score model. The inverse, transforming a mean opinion model into an individual score model is not possible without a very lossy pseudoinverse.

The drawbacks of modelling an individual opinion score are twofold. The first one is that it requires and uses more data and parameters, since the model will be trained on many raters per utterance instead of a single mean score. For more interesting models, individual raters might be modelled to learn their biases, which increases the number of parameters in the model. The other drawback is that the user must aggregate the individual scores to compute a MOS or median score, which requires some additional design and computation.

B. BAYESIAN MODELS

1) PARAMETER UNCERTAINTY

In some frameworks, the model parameters θ, such as mean $\hat{\mu}$ and variance $\hat{\sigma^2}$ are fit without a notion of uncertainty about each parameter. In other words, a point estimate $\hat{\mu}$ of a MOS implies that the model does not indicate a degree of uncertainty or expected error between $\hat{\mu}$ and the true μ. This missing uncertainty is not to be confused with the observed sample variance $\hat{\sigma}^2$ which is also a point estimate with undefined uncertainty for the true variance σ^2. Other frameworks consider the standard error of the sample mean and variance by assuming a given distribution that may or may not match the data. Point estimation of MOS is popular in existing frameworks (e.g. POLQA, and PESQ). Well-calibrated uncertainty is a useful quantity for the user. For example, the user may want to know the bounds of MOS, that is, what MOS value is unlikely to be exceeded for a given utterance for a specified quantile. Additionally, these point estimate models often fit a mapping function by minimizing mean squared error, which means that outliers are penalized. This causes the distribution of the predicted MOS to be under-dispersed when compared to the observed distribution of MOS. That is, extreme MOS values closer to 1 or 5 will be seen less often in the predictions than in the real data. This

compression effect of the true distribution into the narrower predicted distribution will increase as the prediction error increases for point estimate models. Alternatives to models fit with MSE include maximum likelihood or quantile regression [27], which are useful and practical solutions, but have issues with the observed variance problem described below.

When subjective tests are performed, it is common to see the results reported with the sample mean $\hat{\mu}$ specified along with a 95% or 99% symmetric confidence interval $[\hat{\mu} - c\hat{\sigma}, \hat{\mu} + c\hat{\sigma}]$ where $\hat{\sigma}$ is the sample standard deviation of the opinion score and c is a constant. This provides a basic uncertainty estimate about the MOS, but again, this uncertainty assumes that $\hat{\mu}$ and $\hat{\sigma}$ have no error. Additionally, for the discrete 5-point scale used in subjective testing there are problematic properties of this formulation of uncertainty. For example, the symmetry of the confidence interval is not appropriate near the extremes of '1' or '5'. It can also be seen that the confidence interval is difficult to use with a smaller number of ratings, because the variance becomes less reliable. For example, if the study has only 5 raters and they all rate a '5', the variance will be zero, and the confidence interval will also be zero. This problem of variance estimation still exists if the raters give different scores with non-zero variance, although it is less obvious.

Bayesian models resolve the issues seen in the point estimate models and confidence interval analysis. In the Bayesian framework, the appropriate model will provide uncertainty estimates that are tailored to the 5-point outcome space by looking at the distribution of posterior samples. This means they will produce asymmetric credible intervals if the data requires it. The model will also start with a prior selected by the system designer that contains a reasonable distribution for each parameter (with non-zero variance), which handles the problem of sampling limited data where the sample variance does not capture the true variance.

2) ENTROPIC RATIONALE FOR LANGUAGE PREDICTORS

Next, we consider the basis in Bayesian models that suggests that adding language as a predictor will produce more accurate estimates. Given observed scores and any collection of observed features and latent (e.g. learnable) parameters, Bayes' formula provides a description of the uncertainty of each opinion score value as

$$P(score|features) = \frac{P(features|score)P(score)}{P(features)}. \quad (1)$$

Note that this can also be written using the joint distribution on the right hand side as

$$P(score|features) = \frac{P(features, score)}{P(features)}. \quad (2)$$

By modelling the joint distribution of features and scores the probability of each score can be inferred. A Bayesian model is able to model a joint distribution of the features (including features that are parameters and hyperparameters), and is therefore able to give a probability estimate that is precise to the extent that the joint distribution of features and scores has a low entropy.

Previous studies have shown that there is an effect of language on scores [5], [24], so it seems reasonable to infer that score and language are not unconditionally independent, and that the mutual information $I(S; L)$ between score and language is positive. However, score and language may be conditionally independent given the input features (which may contain language information). It follows then that models that only use features with signal-level information, without information about the language will have a strictly higher entropy unless the score is conditionally independent of them given the signal-level features, since for variables S as score, F as some chosen set of features, and L as language,

$$H(S|F) = H(S|F, L) + I(S; L) \quad (3)$$

from which it follows that

$$H(S|F, L) \leq H(S|F), \quad (4)$$

with equality when the features are chosen such that scores are conditionally independent of language information given these features. A higher entropy in score conditional on features that do not contain language information means that the accuracy of the predictions should be worse, because the lowest error a model conditional on these features can achieve will be higher due to the increased uncertainty.

The language information may be present in fine-grained signal features, but typically for opinion score estimation, the signal features are coarse (e.g. a one-dimensional scalar representation of similarity between the degraded and reference signals), and furthermore, the signal only contains the language of the utterance and not necessarily that of the rater (although for the purposes of this paper the main focus is on native language testing). The simplest way to obtain conditional independence between score and language is to add information about the language to the feature set. Alternatively, depending on the causal assumption (described in subsection III-D) that other variables fully contain the language information, such as rater identifiers, it can be sufficient to simply include information about each individual rater without pooling them by language, although including both may be desirable for other reasons.

C. MODEL OUTCOMES

An individual rater score is represented by a discrete value from 1 to 5. These values are ordered by perceived quality. Other Bayesian models have used normal distributions to model MOS [5]. A normal distribution does not model the boundaries at 1 and 5, and further, is continuous, while individual rater scores are discrete. Another work uses multinomial distributions to model histograms of scores for each utterance [20]. The multinomial model has the advantage of capturing a distribution of each of the five scores discretely per utterance, but is not able to model an individual rater's bias over multiple utterances, which is necessary for the experiments involving individual raters. There are several

options for the model that are reasonable given the individual rater problem. Here we consider three different types of models based on their outcomes.

1) ORDERED CATEGORICAL OUTCOMES

The ordered categorical distribution, also known as the ordered logit or ordered logistic distribution, describes an ordered categorical (discrete) variable (such as an opinion score that has categories such as 'bad', 'poor', 'fair', 'good', and 'excellent' with a notion of order). Given N categories, there will be $N-1$ log-cumulative odds $\kappa_1, \kappa_2, \ldots, \kappa_{N-1}$, from which the linear categorical probabilities $p_1, p_2, \ldots, p_N$ can be derived via softmax. Each individual logit is made cumulative by using the probability that the outcome is less than a given cutpoint n, and is given as

$$\begin{aligned} \kappa_k &= log \frac{Pr(y_i \leq k)}{1 - Pr(y_i \leq k)} \\ &= \alpha_k - \phi_i, \end{aligned} \tag{5}$$

where k is the category index and i is the observation, and ϕ_i is any model based on predictors x, such as a simple linear model with one parameter β:

$$\phi_i = \beta x_i. \tag{6}$$

Additional (possibly non-linear) terms will result from adding more predictors. For the opinion score dataset, the features and model of ϕ depends on language, individual rater, and signal similarity and is discussed in section III-E.

2) TRUNCATED NORMAL OUTCOMES

The normal distribution is not bounded, which can be a problem for opinion scores. Rounding the outcome of a normal distribution to the opinion score range can produce undesirable results. An alternative to the ordered logistic distribution that is appropriate for the opinion score problem is the truncated normal distribution, which specifies a PDF that has support entirely within a desired range. This is similar to the normal distribution, but has a lower and upper bound (which would be 1 and 5 for the MOS problem) that asymmetrically truncates the outcome. The probabilities after truncation are normalized by the truncated density so that the resulting PDF sums to 1.0. The main advantage of the truncated normal distribution over the ordered logistic is that it is difficult to formulate an ordered logistic distribution that allows for regression on more than the ϕ parameter. That is, the cutpoint parameters that control the thresholds of each category are typically not indexed per group (e.g. all raters share the same cutpoints, and only differ by their individual ϕ offset). The truncated normal should provide a more flexible model for opinion score estimation, since some raters may have more or less variance but the same mean. The disadvantage is that the truncated normal distribution outputs continuous scalars, which will not match the discrete nature of the opinion scores. As a result, the truncated normal is more useful for applications involving prediction as opposed to simulation.

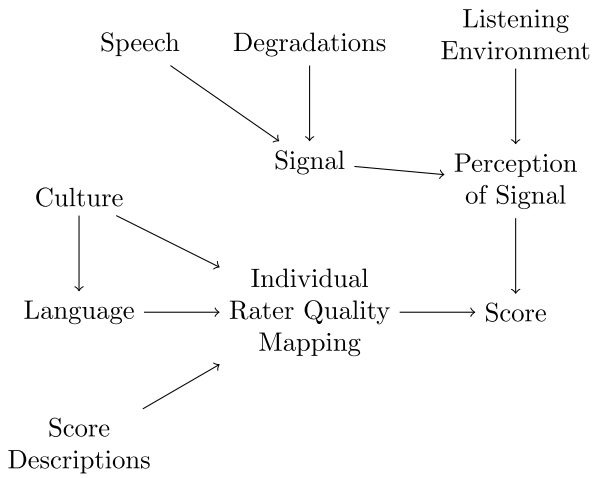


FIGURE 1. A plausible causal DAG for individual ratings. The acoustic properties of the test signal are only a part of the process that determines the score.

3) MOS OUTCOMES

A third option is to model the mean of the opinion score per utterance directly. Here, the data is aggregated before the model is fit, so there is no concept of individual rater. However it is still possible to use language predictors with this type of model. The formulation of this model is very similar to the truncated normal outcome model of individual scores, using the utterance MOS instead of individual scores.

D. CAUSAL MODEL OF OPINION SCORE

For pure prediction problems, it is not strictly necessary to have a thorough understanding of the causal model when designing a regression model. That is, the system designer will generally achieve a more accurate prediction by adding as many predictors as possible, without needing to worry about causation versus correlation, or confounders. Deep learning presents many useful examples of this by using as many input features as possible. However, many popular MOS estimation frameworks use only signal level features, which ignore information about the rater and their environment.

It may be useful to consider a plausible causal DAG such as the one in Figure 1 to entertain potential non-signal predictors. In the DAG, several variables independent of the speech signal are considered. 'Culture' contains many attributes, and is an unobserved variable, but since culture usually determines (native) language, language may serve as a proxy for culture, which may include rating tendencies. The 'Environment' variable pertains to the listening environment that the rater conducts the test in, and contains many potentially unobserved attributes, such as the listening equipment and presentation order of the signal (which may cause listening fatigue), and even the weather during the test. The score description labels, which are text in the native language presenting along with each rating level (e.g. 'poor' in English, 'warui' in Japanese, or 'médiocre' in French for

the score '2'), also affects the likelihood of a certain score, because the text may imply different qualities in different languages. Furthermore, the choice of labels may imply a nonlinear scale within a single language if the labels are not perceived to be equidistant. The culture and language problem has been studied to a considerable degree, for example, in [23], with some solutions using models specific to a certain language [28], [29].

If unlimited computing capacity is available, it would be desirable to add as many of these predictors to the model to obtain the highest accuracy. But in practice, compute and data are scarce resources. Furthermore, the causal process is noisy, and it is unclear to what extent each feature is actually useful in a predictive model without experimentation. For example, neither culture nor region uniquely determines language. An experiment is needed to see to what extent information about language improves the accuracy of the model, and so on for the other variables.

E. FEATURES AND PARAMETERS

MOS estimators typically will include features that are extracted from the speech waveform. For this problem it is appropriate to use the neurogram similarity index measure (NSIM), which is a 1-dimensional indicator of similarity over all frequency bands and time between the reference and degraded signals. NSIM has shown to be useful for early versions of ViSQOL [30], and the one-dimensional property allows for a relatively simple model with a single parameter related to the signal. Signal predictors which provide more modelling power and less aggregation certainly exist (e.g. multiple frequency band NSIM [31], mel-spectrogram, or WARP-Q [32]). However, the purpose of this study is not to find the most useful signal-level descriptors, but instead to find the effects of features that are external to the signal, such as the individual rater and language bias. To this purpose, it is desirable to keep the signal level features minimally complex. Figure 2 illustrates the features that are used by the Bayesian model as predictors.

As previously mentioned, individual raters have a bias and variance that differs from other raters. A rater identifier feature that is unique for each rater is added to the model to allow it to be aware of this bias. Similarly, language identifiers can be a feature that uniquely identifies the language. Because the raters of a given language forms a group, it is sensible to apply a hierarchical model to share information between individual raters. For the purposes of this study, this 'language' feature will be a laboratory identifier where the native language is used to test, and also encompasses other factors in the entire test environment such as the culture of the laboratory, the listening equipment, and so on. Each rater and language identifier is used as an index variable with normal priors that linearly influence the ϕ offset for the ordered logit model, and an exponential model for NSIM, as was found to be useful in [7]. The prior for ϕ_i can be described for individual observation i, rater j, and language k,

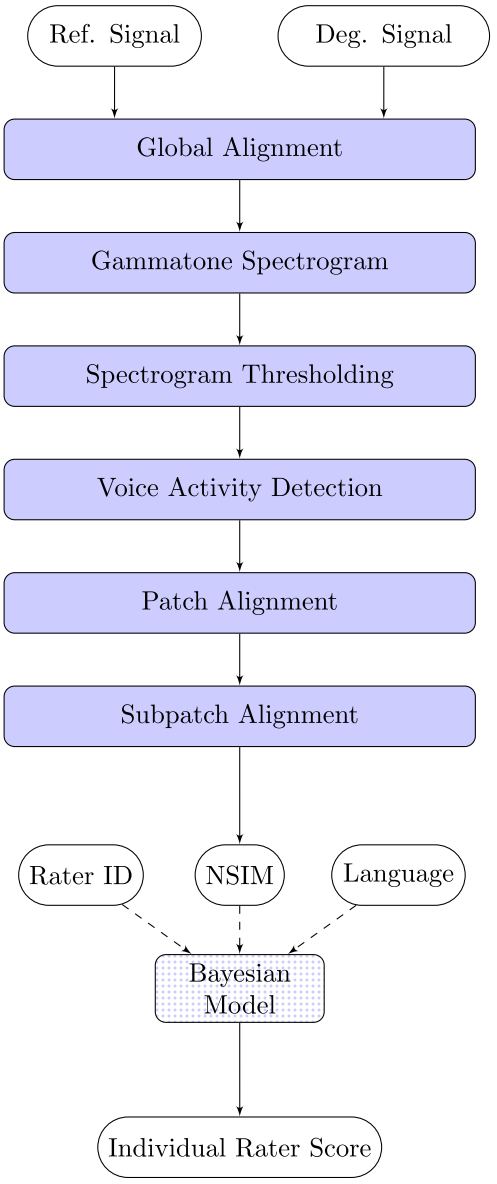


FIGURE 2. A system diagram showing the features that are used as predictors in the model. The dashed arrows represent optional predictors.

and NSIM observation x_i as

$$\begin{aligned} \phi_i &= \alpha_j + \gamma \\ \alpha_j &= \text{Normal}(\mu_k, z) \\ \mu_k &= \text{Normal}(a, b) \\ \gamma &= e^{\beta(x_i - \theta)} \\ \beta &= \text{Normal}(c, d) \\ \theta &= \text{Normal}(e, f), \end{aligned} \tag{7}$$

with user-defined constants a, b, c, d, e, f (to be found with prior predictive checking). Note that the priors with subscripts denote that there is one prior for each item in the group, e.g. the μ_k indicates multiple Normal priors, one for each of the k languages, so the model will fit each item of the group separately. The truncated normal model is formulated

similarly, with ϕ_i being used as the mean for the truncated normal distribution, with an additional prior for variance.

F. COMPUTING MOS FROM INDIVIDUAL SCORE MODELS

An individual score model as described above outputs a discrete score for each utterance and rater. MOS is often used as the mean aggregated over utterances, or a collection of utterances within a certain test condition. Given an individual rater's score probability $p_r(s|x)$, the true MOS over a set of utterances X and raters R can be computed as

$$\begin{aligned} \text{MOS}(X, R) &= \mathbb{E}[s|X, R] \\ &= \frac{1}{|X||R|} \sum_{x \in X} \sum_{r \in R} \sum_{s \in S} p_r(s|x) s \end{aligned} \tag{8}$$

where s is the score in the set $S = \{1, 2, 3, 4, 5\}$. For example, the utterance MOS uses X with a single element (a single utterance), and the condition MOS uses X with all the utterances in the condition.

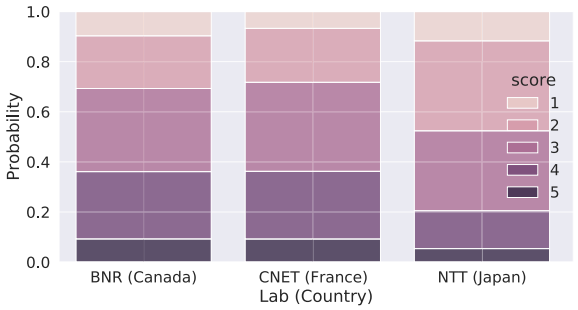
$p_r(s|x)$ is an unknown quantity that must be estimated. In a Bayesian model each posterior sample is a sample from the joint distribution of all the parameters, so multiple samples will produce different likelihoods for the same utterance and rater pair. So in practice estimating $p_r(s|x)$ requires multiple samples. One way to do this is to sample the model's joint distribution many times to obtain the probability by converting the histogram into a probability. However, since we are interested in the expectation over X and R, the process can be further reduced by simply taking the mean of the posterior scores as the estimated MOS. In other words, although the observed data ratings have each listener rate each utterance once, in the posterior samples each listener 'rates' each utterance hundreds or thousands of times, and the MOS can be estimated by taking the average of all samples.

IV. EXPERIMENTS

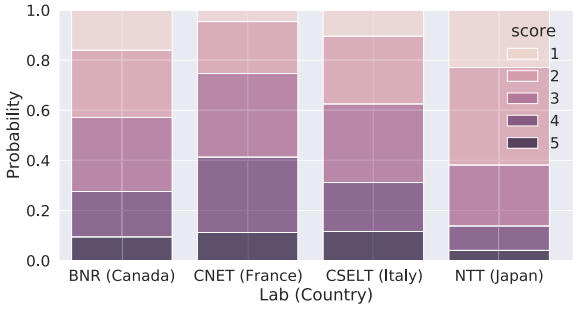
In this section, an appropriate dataset is analyzed, and experiments that use proposed models are presented.

A. DATASET

The ITU-T P. Supplement 23 dataset (experiments 1 and 3) is well-suited for this experiment. It is conducted across four different languages and laboratories with all listeners being native speakers of the language used in the utterance. Additionally, the recording conditions, and the signal processing chain including the pre- and post-processing of the signals are well-documented, and each lab conforms to the shared procedures. Lastly, all of the labs in a given experiment tested the same conditions (e.g. street noise at 6dB SNR), so the results between the labs should be comparable. These properties of the dataset enable this experiment to measure the effect of language. There are 24 listeners in each experiment and laboratory combination, and each listener rates many utterances, with the order of presentation randomized according to 4 different randomization patterns. All of the data for individual raters is recorded in the dataset (i.e., the data is not aggregated into MOS).



(a) Experiment 1



(b) Experiment 3

FIGURE 3. Distribution of individual rater scores in the ITU P Supp. 23 dataset for different languages. Each laboratory conducted the experiment under the same conditions in the native language. 'BNR' is Bell Northern Research in Ottawa and uses English.

1) ANALYSIS OF DISTRIBUTIONS BY LANGUAGE

In the P. Supplement 23 dataset, experiment 1 tests the performance of low bitrate codecs with transmission standards. Experiment 3 tests the effects of channel degradations. It may be useful to consider the distribution of the observed data for both of these experiments.

Figure 3 compares the observed distribution of each rating between labs and experiments. There are remarkable differences between the different languages. Japanese raters tend to rate with very few 'fives', and many 'ones'. French raters are the opposite, being the least likely to rate as 'ones', and the most likely to rate 'fives'. It is expected that the content of the experiment affects the distribution of scores. So while it is expected that the score distributions within a language change in different experiments, it is interesting to note that the distribution preserves some properties, such as the relative biases of the languages.

2) ANALYSIS OF INDIVIDUAL RATER DISTRIBUTIONS

Looking at the individual rater distributions within each language naturally contains all of the information to describe the distribution of the language, since the language data is simply the set of all raters of the language. But it also contains some additional information pertaining to each individual rater's tendencies that is not in the language information alone. In figure 4 it can be seen that there are language-level biases, and within a language that there are individual rater biases. So it appears to be reasonable to construct a model that captures both language and rater information.

B. MODEL SPECIFICATION

Several models are considered to show the effects of different predictors. These models and the features they use are described in table 1. The most basic model 'Baseline', uses no predictors (i.e. no input features) to predict a score. This model will obviously not be able to predict the scores of individual listeners or utterances accurately, but it should be able to model the overall distribution of the input. It serves both as a control that other models can be compared against, as well as to test whether the distribution of observed ratings has a dispersion that is captured by an ordered categorical model with the provided priors.

The most complicated model 'LangRaterNSIM' is the one that uses the features described in III-E that we expect to contain information about the outcome score. More specifically, the predictors are listener and language identifiers (indices) along with a scalar NSIM predictor that indicates signal similarity between the original reference and the degraded utterance that the rater has scored. Additionally, a model called 'Order' with a single predictor uses the logarithm of the presentation order of the utterance to predict the score.

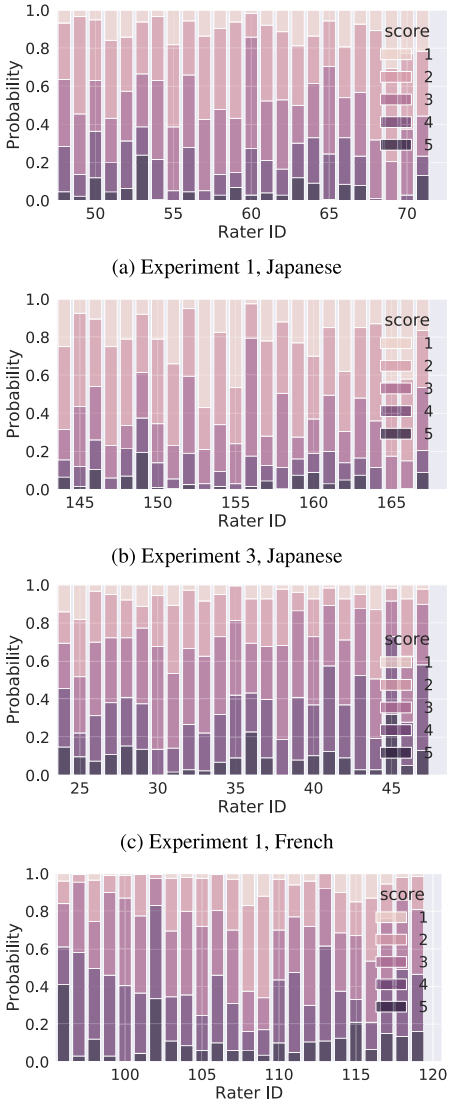
TABLE 1. Models and predictors.

	Predictors			
Name	Language	Rater	NSIM	Order
Baseline				
Order				✓
Lang	✓			
LangRater	✓	✓		
NSIM			✓	
LangNSIM	✓		✓	
LangRaterNSIM	✓	✓	✓	
RaterNSIM		✓	✓	

Additionally, we fit two models (NSIMMOS and LangNSIMMOS) that have the same predictors as NIM and LangNSIM, but are fit to pre-aggregated utterance MOS data directly instead of using individual rater score, and a truncated normal model (LangRaterNSIMTrunc) that has the same predictors as LangRaterNSIM. All models that do not end with either 'MOS' or 'Trunc' are ordered logistic models.

C. MOS RESULTS

We use the method described in section III-F to estimate MOS for each utterance, condition, and lab-specific condition. Table 2 shows the error and correlation coefficients for each model for three types of aggregation. The most common of these in the literature is aggregation by condition, which generally produces the lowest error and highest correlation, followed by aggregation by utterance, which produces a higher error due to the smaller number of samples. To better understand the effects of language, an aggregation by condition within each language is also presented. The models that add language predictors have a relatively large improvement for the aggregation by language and condition (0.562 vs 0.464 RMSE for NSIM vs LangNSIM), and the differentiation of language is evident in figure 7. Additionally, figure 6 visualizes the predictions in joint plots with the



(a) Experiment 1, Japanese

(b) Experiment 3, Japanese

(c) Experiment 1, French

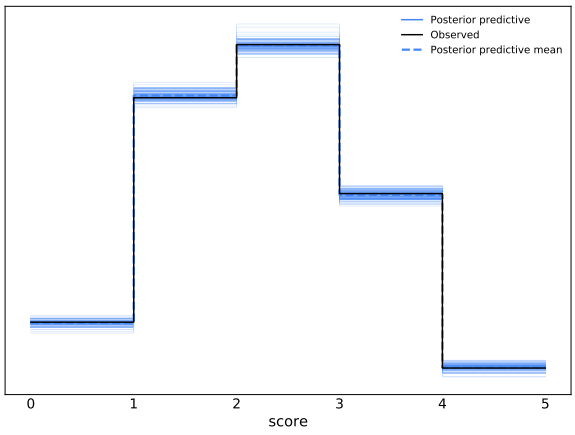
(d) Experiment 3, French

FIGURE 4. Distribution of scores for each of the 24 raters in the ITU P Supp. 23 dataset for Japanese and French amongst the two different experiments. Japanese and French raters are the most different from each other in that they tend to rate low and high respectively. The distributions are different enough to be visually apparent, but have enough variance that there is overlap - the highest rating Japanese rater tends to rate higher than the lowest rating French rater.

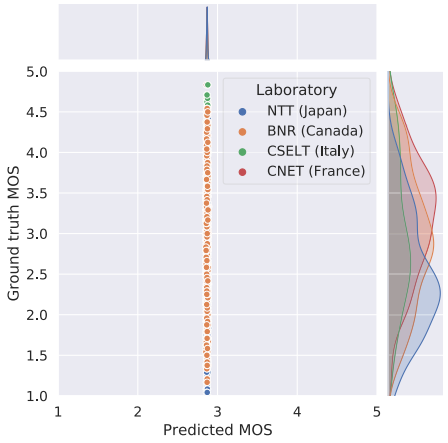
ground truth MOS at the utterance level, where the language effect is also evident.

The 'Baseline' model has no predictors at all and its MOS predictions naturally converge on the MOS over all utterances, which is a value just below 3.0. The distribution of the unaggregated individual score predictions matches the input distribution well. This can be seen in figure 5.

For comparison purposes ViSQOL exponential and PESQ are added as anchors that should align with the Bayesian NSIM model, as they are models that map a single dimension of physical similarity (e.g. mean similarity across all frequency bands and time) to quality that is analogous to the NSIM predictor used by the Bayesian models. POLQA



(a)



(b)

FIGURE 5. The baseline model is able to model the distribution of scores in the posterior samples (a). The individual predictions have no specific information, so although individual predictions span the full score range, the MOS of any individual utterance converges on the global MOS as shown in the joint plot (b).

is added as a state of the art model that is known to perform especially well on this dataset. The Bayesian NSIM model performs on par or slightly better for RMSE than PESQ and ViSQOL exponential [7], indicating that it makes reasonable usage of the information in the signal predictor. It is worth noting that PESQ has a relatively high correlation for the condition aggregated cases even when its error is higher. As mentioned earlier, the purpose of this study is not to compare a simple Bayesian model with state of the art models that have complex predictors (e.g. multi-dimensional predictors with similarity for multiple frequency bands), but to exploit the simplicity of the Bayesian model to test the marginal benefit of adding language features on top of signal features. For this reason the POLQA scores are not directly comparable, and it is not surprising to see that POLQA has lower RMSE across the board.

D. MODEL VALIDATION AND COMPARISON

Model design, validation, and comparison of Bayesian models goes beyond looking at error and correlation. It involves an interactive process that is facilitated with prior predictive checks, posterior predictive checks, and verification that the posterior samples are useful. For example, inspecting the posterior samples against the observed appears to be reasonable at the global outcome level in figure 8a. These posterior samples can also be used to create a measure of uncertainty if the models are reasonably calibrated by confirming that for a certain quantile, approximately that many samples are underestimates (e.g. the median quantile should have half of the posterior estimates above the observed median.) The estimates at the global and language level appear to be reasonably calibrated, but this is not so for all individual raters, so the uncertainty should only be useful when aggregating over utterances, as is the practice with MOS.

Bayesian models are typically fit on the entire dataset, unlike deep learning models that split the data into a train and a test set. There are several reasons for this. Bayesian models are always probabilistic models, and other metrics that rely on this property can be used to check that the model is not invalid and that it is able to predict out of sample data, such LPPD, PSIS, WAIC, $\hat{r}$, and visual inspection of the chains and the posterior predictive distribution. It is common to compute estimates of 'leave one out' cross-validation (LOOCV) to check for out of sample predictive accuracy without dividing the dataset into train and test splits. PSIS and WAIC are popular estimates for this purpose. These metrics also measure predictive power, i.e. how accurately it will predict an out of sample observation.

Since some of the models in this experiment have different outcomes (i.e. MOS models versus individual score models), not all models are suitable for relative comparisons of WAIC and PSIS. For select models where the comparison is sensible, table 3 shows the values for WAIC and PSIS, which converge on the same values within a decimal point. Here, higher values indicate more accurate out of sample predictions. The ranking matches the original models RMSE rankings. If the two models have the same predictive power the model with a language predictor is preferable to allow for predicting language effects.

The statistical significance of these results should be discussed. ITU-T Rec. P. 1401 [33] provides statistical tests for comparing the significance of model differences in RMSE. Under these tests, the RMSE difference is significant between 'Baseline' and 'Lang' ($p = .0395$), between 'Lang' and 'NSIM'($p = 8.01\text{e-}8$), and between 'NSIM' and 'LangNSIM'($p = 2.58\text{e-}5$), but not between 'LangNSIM', and 'LangRaterNSIM' ($p = 0.254$). This is more evidence in favor of language being a useful predictor on top of a signal predictor like NSIM, and that rater information may be useful, but not significant under this test. However, the significance of predictive power can also be looked at on the outcome scale of individual rater scores instead of aggregating the raters into a single value. For this purpose, the PSIS/WAIC analysis also provides the function of a statistical test for the significance of the out of sample predictive performance. It is also important to point out the relatively small standard errors in table 3

TABLE 2. Model comparison.

	Utterance			Language+Condition			Condition		
Name	RMSE	Pearson	Spearman	RMSE	Pearson	Spearman	RMSE	Pearson	Spearman
Baseline	0.829	0.043	0.035	0.778	0.106	0.090	0.705	0.151	0.123
Order	0.829	-.0326	-0.033	0.779	-0.029	-0.035	0.706	-0.039	-0.033
Lang	0.790	0.303	0.295	0.734	0.333	0.321	0.707	-0.050	-0.023
LangRater	0.785	0.321	0.319	0.729	0.362	0.351	0.704	0.105	0.136
NSIM	0.684	0.568	0.551	0.562	0.717	0.700	0.431	0.858	0.856
NSIMMOS	0.681	0.570	0.554	0.560	0.712	0.699	0.427	0.858	0.856
LangNSIMMOS	0.613	0.674	0.664	0.457	0.821	0.812	0.404	0.850	0.840
LangNSIM	0.612	0.670	0.659	0.464	0.822	0.813	0.415	0.851	0.843
RaterNSIM	0.603	0.687	0.678	0.446	0.840	0.839	0.407	0.861	0.853
LangRaterNSIMTrunc	0.603	0.686	0.677	0.444	0.839	0.836	0.405	0.858	0.859
LangRaterNSIM	0.601	0.689	0.679	0.446	0.839	0.837	0.409	0.859	0.852
ViSQOL Exponential	0.740	0.498	0.488	0.652	0.685	0.676	0.562	0.838	0.836
PESQ	0.665	0.771	0.760	0.607	0.809	0.804	0.486	0.910	0.935
POLQA SWB	0.524	0.783	0.756	0.430	0.837	0.808	0.182	0.970	0.972

TABLE 3. Predictive power and significance.

Name	WAIC/PSIS(LOO-CV)	Standard Error
Baseline	-48109	74.30
Lang	-47217	82.29
NSIM	-45085	90.51
LangNSIM	-43717	97.56
RaterNSIM	-40487	110.07
LangRaterNSIM	-40485	110.49

due to the large amount of data. This shows that most of the models do not overlap within three standard errors, and that there is a real benefit to predictive power from including each of the features in this order. The exception is the overlapping LangRaterNSIM and LangRaterNSIM, which is expected due to the rater data fully containing the language data.

Another difference from deep learning is that in Bayesian models, there are fewer parameters which make overfitting less of a risk, and models that have too many parameters are often non-identifiable (which would fail the $\hat{r}$ test, or would be so specific as to not be useful (e.g. a separate model for each observation). The hyperparameters of a Bayesian model are the distribution parameters of the root level priors, and these are typically set at model creation time, or interactively, by looking at the prior predictive distribution (which does not involve the data). Some of the more accurate models in this experiment have very few parameters (LangNSIM only has 10 parameters: 4 for the cutpoints, 4 for the language offsets, and 2 for NSIM slope and intercept).

The models were fitted using Hamiltonian Monte Carlo (HMC) which efficiently samples the posterior distribution with a No-U-Turn sampler [34] using the TensorFlow Probability framework [35]. Multiple chains are used in HMC, and the typical check of involves visual inspection of the chains as well as verifying that $\hat{r}$ values are reasonable (i.e. near 1.0) to check that the chains converge and do not get stuck on some values, which would indicate that the posterior was not well-explored. For the models studied in this paper these checks have passed, and the posterior distributions of the parameters and their chains appear healthy as shown for the LangRaterNSIM model in figure 9.

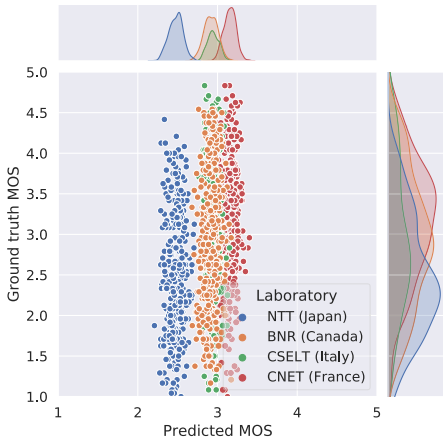
V. DISCUSSION

A. EFFECT OF LANGUAGE AND RATERS ON OPINION SCORES

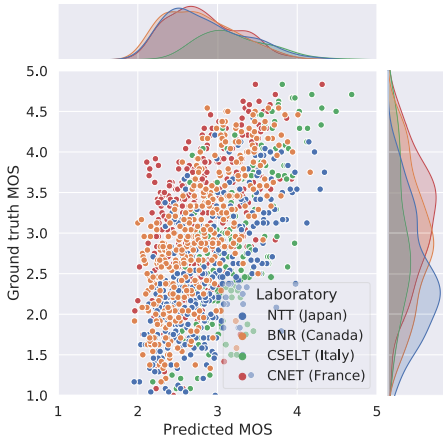
In table 2 it can be seen that adding language predictors improved RMSE over a baseline model with no predictors (from 0.829 to 0.790) and correlation (from 0.043 to 0.303). Adding rater information improves it further. Furthermore, the improvement when adding language and rater information is still significant on models that have signal level predictors (from 0.684 to 0.601 RMSE, 0.568 to 0.689 Pearson), which indicates that coarse signal predictors (of the type used in typical MOS estimation tools) cannot tell the full story about opinion scores. This is consistent with the observed opinion scores grouped by language in the data that compare the same degradations in figure 3.

The findings of this study show that a correlation between language and score exists, which agrees with previous work, but also shows the increase in modeling power when including language as a predictor. It has not been concluded whether the differences in scores between languages are cultural responses or related to perceptual quality. The causes behind the bias are related to general item response theory and psychology, and requires more complicated experiments such as cross language studies with bilingual listeners to resolve, as well as studies that consider ratings that are completely unrelated to speech quality (e.g. restaurant ratings).

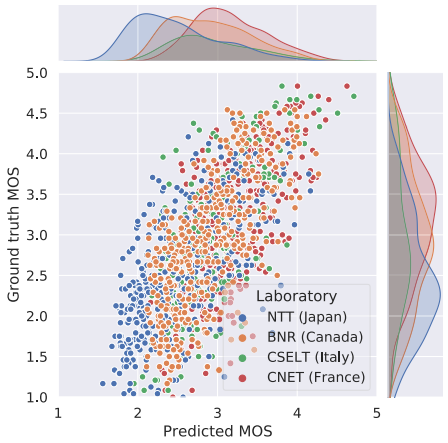
For example, the finding of a lower bias in Japanese scores does not reveal whether the 'true' quality of Japanese speech is lower, or that Japanese raters tend to rate a given quality lower than other raters. That question remains unanswerable with the current study. It is plausible that the distribution of phonemes in the language influences the score, just as it is plausible that there is a culture to rate things lower. In other words, the subjective test only measures the categorical ratings from 1 to 5, and how a different quality level is mapped to these numbers is unobserved. However, leaving the elusive unobserved 'quality' aside, it is possible to infer equivalencies between scores for different languages



(a) LangRater

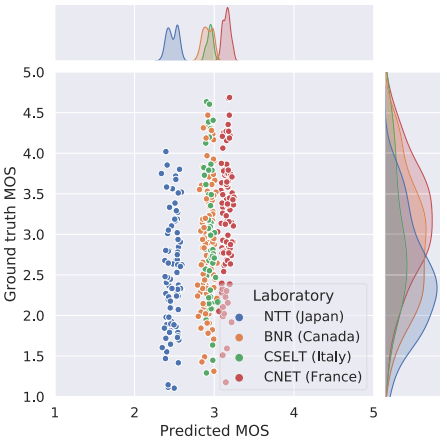


(b) NSIM

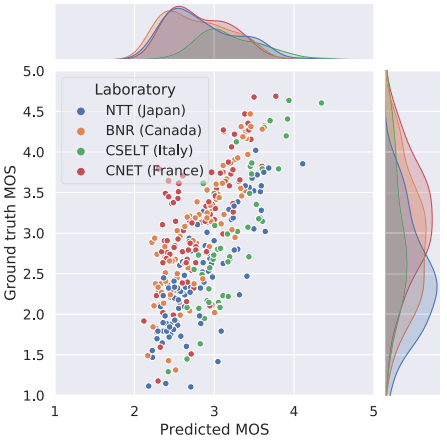


(c) LangRaterNSIM

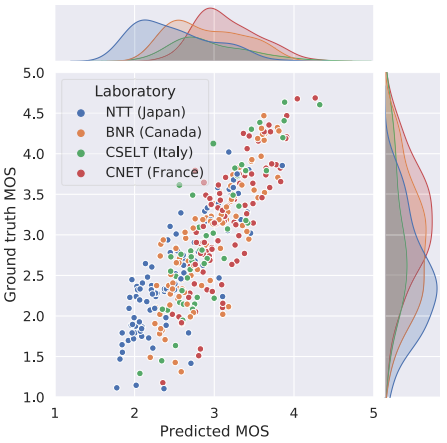
FIGURE 6. Joint plots of per-utterance mean opinion scores for select ordered logistic models. Because the LangRater model has no signal predictor, it can only estimate language and rater means. The NSIM model has only a signal-level predictor and is not able to capture the differences in languages. LangRaterNSIM with signal, language and rater predictors improve on the NSIM model differing modes of each lab, showing the marginal improvement over NSIM.



(a) LangRater



(b) NSIM

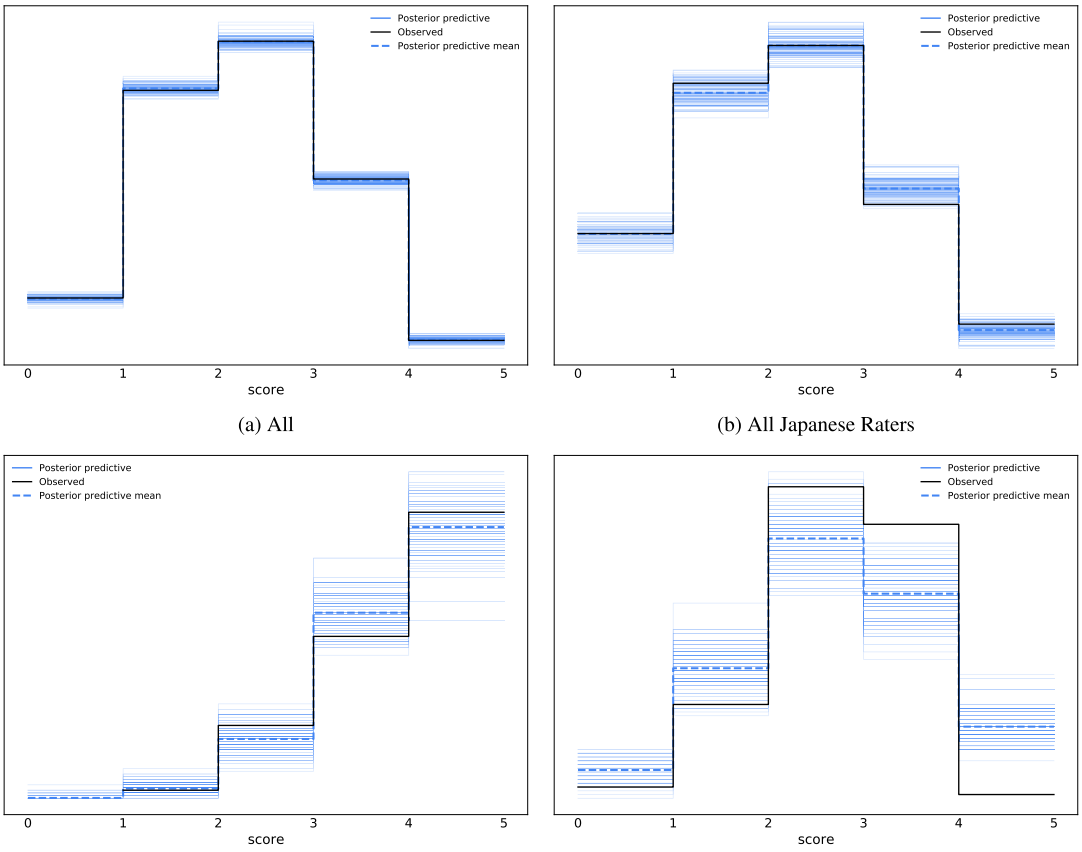


(c) LangRaterNSIM

FIGURE 7. Joint plots of per-language-and-condition mean opinion scores for select ordered logistic models. Aggregating over conditions, which are a larger group than utterances, reduces the error and increases the correlation. The effect of language and rater can still be seen in the marginal distributions as in figure 6.

and signals. For example, for a certain kind of degradation, the expected Italian score is X, and for Canadians it is Y.

This does suggest that opinion scores should not be compared absolutely between different languages. The ITU-T specifications for P.800 listening tests [2] also gives strict recommendations for comparing MOS that excludes the



(a) All

(b) All Japanese Raters

(c) Individual Japanese Rater A

(d) Individual Japanese Rater B

FIGURE 8. Posterior samples from the LangRaterNSIM ordered logistic model compared to the observed scores for the overall data and different subgroups. The posterior overlaps well for the global (a) and language group (b), and reasonably for (c). The individual rater B (d) is an imperfect fit, presumably because the model uses shared cutpoints for all raters and is not able to increase the mean without increasing the likelihood of the '5' score.

cross-language case. It should be noted that while the model that was fit in this study has the ability to answer cross-language counterfactual questions such as 'what score would an Italian rater assign to this English utterance?'. However, there are zero occurrences of cross-language ratings in the particular dataset used in our experiments, so it should only be used with the due amount of apprehension.

Both individual rater models and MOS models can be made language aware, as has been done in the experiment. The findings with improved accuracy for language based models suggest that language predictors should be added to a scoring model, unless there is only a single language present in the data, and the model will never be used to predict scores for other languages. Since language is being used here as a proxy for culture, it may be useful to include other cultural covariates such as region and year if there is variation within languages in the data. To test the hypotheses of this experiment it was sufficient to use one laboratory per language at a single point in time. This also means that the resultant models may not generalize to other populations within the same language. As mentioned previously, the purpose of this paper is not to obtain the most generally useful and most accurate model, but to answer the questions about the effect of language and raters in opinion score data.

The RaterNSIM model performed similarly to the LangRaterNSIM, which has additional parameters for language. This is consistent with the causal model DAG given in figure 1, because the individual rater blocks the path between language and score, so score and language are conditionally independent given the rater information. However, the advantage of the model with language parameters is that it is straightforward to simulate or answer counterfactual questions about new raters in the same language or to 'translate' raters into other languages.

B. INDIVIDUAL SCORE VERSUS MOS MODELS

For the purposes of this paper, we are only concerned with identifying the distribution of each rater's scores with a model that is given information about each rater, by also fitting a distribution of raters. Having fit such a model, it is then possible to predict the behavior of each rater with a higher accuracy. Two types of individual rater models were considered: the ordered logistic model and the truncated normal

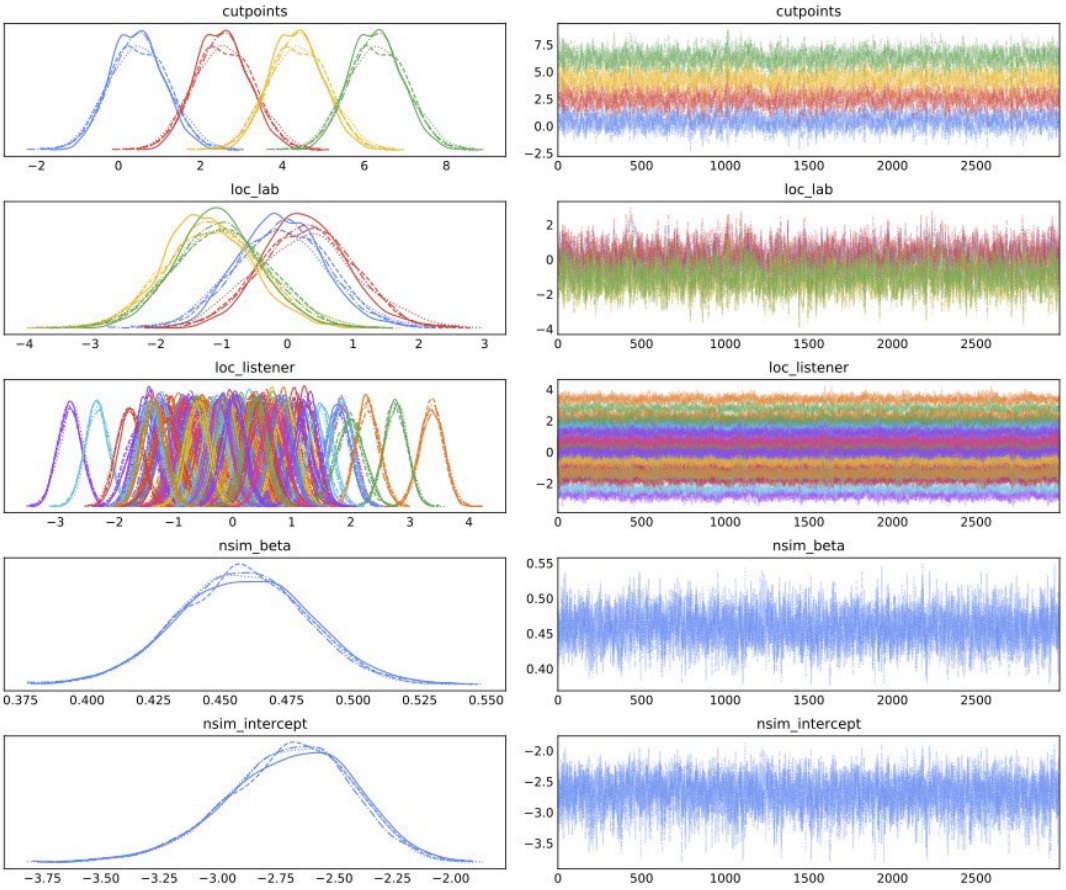


FIGURE 9. Traceplot of parameters for LangRaterNSIM model. The smoothed posterior samples form distributions for each indexed parameter, shown on the left in different colors (e.g. each language/laboratory is a different color) with each line-style (e.g. dashed, solid) representing the chain index. The Gaussian noise-like patterns on the right indicate the posterior for each chain and index is being explored in a healthy manner. The overlapping chains of the same color show that the model is identifiable.

model. The truncated normal model achieved a lower error, presumably because in the truncated normal, the variance and mean are modeled as parameters for each rater, but in the ordered logistic model this is not the case because of the cutpoint formulation (only a logit offset, or 'location' parameter is modeled at the individual level). The ordered logistic distribution may still be useful, depending on the application. For example, unlike the truncated normal distribution, the ordered logistic distribution captures the discrete nature of categorical opinion scores, which may be desirable if the goal is to generate simulated data for individual raters.

The experiment compared models fit to MOS data as well as individual rater data. The experiments found that individual rater models are able to match and exceed the accuracy of MOS models because of the extra information they have, although the difference is not very large (.613 vs .601 utterance RMSE). The variance of individual raters is an important piece of information that is discarded in most MOS models, but is accounted for in individual rater models. Additionally, the bias or expected rating for individual raters is discarded by MOS models. This means that the MOS model will spuriously attribute individual rater bias to other predictors, increasing the error, although for this experiment it was not significant (the MOS models performed similarly to the individual rater models). For example, suppose there are two utterances of equivalent quality, and a rater that consistently gives low scores. If the rater rates the first utterance, but does not rate the other, it should be expected that the MOS for the first utterance will be lower than the second. The MOS model, which does not have access to individual rater information, will not be able to recognize that the first utterance's lower score is due to the pessimistic rater's presence and will instead will attribute it to the signal predictors, or produce a wider uncertainty on all utterances of this quality. In contrast, the individual score model will handle this case by attributing the difference in scores to the pessimistic rater.

If a distribution of the individual rater is modeled, the model can answer questions that a MOS model cannot, such as 'what would the median rater score this utterance?'. Depending on the application, the median rater's expected score may be desirable over the MOS because the overall variance will be reduced. The individual rater model can be used to post-screen outliers even after being fit on them by

sampling raters from a truncated distribution that excludes the extreme values (e.g. raters within the 5 to 95 percentile).

The drawbacks of the individual score model must be considered. To compute a mean score with an individual score model, multiple samples must be taken and aggregated. This aggregation has computational cost, but the time it takes is relatively quick compared to the time it takes to fit a model.

Lastly, this experiment was concerned with speech data, but the findings may be relevant to non-speech audio as well. The causal DAG in figure 1 proposes that culture and language of the rater, which is independent of the content of the audio signal, affect scores. In this case the individual rater model has the potential to become more valuable because a given test signal may be more readily listened to by people from different languages and cultures, especially with online testing. An experiment to verify this would be prudent.

VI. CONCLUSION

This paper has shown that language-aware models provide a significant improvement over models that do not consider language, and that models of individual scores are able to match and exceed MOS models (in all aspects other than computational cost) while providing additional functionality. The experimental results validate the theory for these arguments.

The findings were over a single dataset, ITU-T P. Supplement 23, which was chosen because of the breadth of data it contains and the thorough process used to create it. However, the dataset is over 20 years old, and enough time has passed that subjective quality may have changed. Additionally, subjective tests are now conducted in a wider variety of environments, such as crowd-sourced tests at home conducted over the internet. It would be interesting future work to re-evaluate the same data in new tests in the same regions, to see how raters have changed since the creation of the P. Supp 23 data was created.

Real world subjective test data does not often have equally balanced experimental conditions over multiple languages as P. Supp. 23 does. However, one of the strengths of individual score Bayesian models is that they are able to handle unbalanced data (e.g. different numbers of listeners for each utterance). In this case, since the utterance quality might be different between languages, score comparisons between the languages may not be useful to look at, but the improvement in accuracy due to the language and rater metadata predictors can be measured. Based on the current results, further studies and applications for this type of data seem like a good next step.

Lastly, the findings about language and individual scores imply a causal model that should apply to non-Bayesian models. For example, deep learning models of speech quality could consider taking into account individual ratings and language metadata in the feature set. Bayesian models are relatively difficult to fit as the data or parameter size grows to very large sizes (because typically the probabilities are computed over all the data), while deep learning can use mini-batches to handle virtually unlimited amounts of data. Given the recent advances in deep learning, it may be interesting to evaluate a model with many input features including language and individual rater score.

REFERENCES

[1] W. B. Kleijn, A. Storus, M. Chinen, T. Denton, F. S. Lim, A. Luebs, J. Skoglund, and H. Yeh, "Generative speech coding with predictive variance regularization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6478–6482.

[2] *Methods for Subjective Determination of Transmission Quality*, document ITU-T Rec. P.800, International Telecommunication Union, Geneva, Switzerland, 1996.

[3] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6493–6497.

[4] N. Osaka, K. Kakehi, S. Iai, and N. Kitawaki, "A model for evaluating talker echo and sidetone in a telephone transmission network," *IEEE Trans. Commun.*, vol. 40, no. 11, pp. 1684–1692, Nov. 1992.

[5] I. Mossavat, P. N. Petkov, W. B. Kleijn, and O. Amft, "A hierarchical Bayesian approach to modeling heterogeneity in speech quality assessment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 136–146, Jan. 2012.

[6] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOL: The virtual speech quality objective listener," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2012 pp. 1–4.

[7] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," in *Proc. 2th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2020, pp. 1–6.

[8] M. Chinen, J. Skoglund, and A. Hines, "Speech quality estimation with deep lattice networks," *J. Acoust. Soc. Amer.*, vol. 149, no. 6, pp. 3851–3861, Jun. 2021.

[9] *Perceptual Objective Listening Quality Assessment*, document ITU-T Rec. P.863, International Telecommunication Union, Geneva, Switzerland, 2018.

[10] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement Part I," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384, 2013.

[11] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, document ITU-T Rec P.862, International Telecommunication Union, Geneva, Switzerland, 2001.

[12] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 749–752.

[13] J. A. Bergstra and C. Middelburg, *The E-Model: A Computational Model for Use in Transmission Planning*, document ITU-T Rec. G.107, 2003.

[14] L. Ding and R. A. Goubran, "Speech quality prediction in VoIP using the extended E-model," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2003, pp. 3974–3978.

[15] *Coded-Speech Database*, document ITU-T P.23, International Telecommunication Union, Geneva, Switzerland, 1998.

[16] H. Gamper, C. K. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 85–89.

[17] J. Serrà, J. Pons, and S. Pascual, "SESQA: Semi-supervised learning for speech quality assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 381–385.

[18] A. Ragano, E. Benetos, and A. Hines, "More for less: Non-intrusive speech quality assessment with limited annotations," in *Proc. 13th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2021, pp. 103–108.

[19] R. Z. Jimenez, L. F. Gallardo, and S. Moller, "Influence of number of stimuli for subjective speech quality assessment in crowdsourcing," in *Proc. 10th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2018, pp. 1–6.

[20] M. Seufert, "Fundamental advantages of considering quality of experience distributions over mean opinion scores," in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6.
[21] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, "QoE beyond the MOS: An in-depth look at QoE via better metrics and their relation to MOS," *Qual. User Exper.*, vol. 1, no. 1, pp. 1–23, Dec. 2016.
[22] J. Nawala, L. Janowski, B. Cmiel, and K. Rusek, "Describing subjective experiment consistency by p-value P–P plot," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 852–861, doi: 10.1145/3394171.3413749.
[23] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests—A review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451, 2008.
[24] Z. Cai, N. Kitawaki, T. Yamada, and S. Makino, "Comparison of MOS evaluation characteristics for Chinese, Japanese, and English in IP telephony," in *Proc. 4th Int. Universal Commun. Symp.*, Oct. 2010, pp. 112–115.
[25] M. Nakayama and Y. Wan, "Same sushi, different impressions: A cross-cultural analysis of yelp reviews," *Inf. Technol. Tourism*, vol. 21, no. 2, pp. 181–207, Jun. 2019.
[26] G. Chen and V. Parsa, "Loudness pattern-based speech quality evaluation using Bayesian modeling and Markov Chain Monte Carlo methods," *J. Acoust. Soc. Amer.*, vol. 121, no. 2, pp. EL77–EL83, Feb. 2007.
[27] R. Koenker and G. Bassett, Jr., "Regression quantiles," *Econometrica, J. Econ. Soc.*, vol. 46, no. 1, pp. 33–50, 1978.
[28] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," *IEEE Commun. Mag.*, vol. 42, no. 7, pp. 28–34, Jul. 2004.
[29] A. Takahashi, H. Yoshino, and N. Kitawaki, "Quality assessment methodologies for IP-telephony services," *IEICE Trans. Commun.*, vol. 88, pp. 863–874, May 2005.
[30] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Commun.*, vol. 54, no. 2, pp. 306–320, Feb. 2012.
[31] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *J. Acoust. Soc. Amer.*, vol. 137, no. 6, pp. EL449–EL455, Jun. 2015.
[32] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "WARP-Q: Quality prediction for generative neural speech codecs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 401–405.
[33] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, document ITU-T Rec. P.1401, International Telecommunication Union, Geneva, Switzerland, Jun. 2020, pp. 401–405.
[34] M. D. Homan and A. Gelman, "The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *J. Mach. Learn. Res.*, vol. 15, no.1, pp. 1593–1623, Jan. 2014.
[35] J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. Hoffman, and R. A. Saurous, "Tensorflow distributions," *Probabilistic Program. Lang., Semantics, Syst. (PPS)*, Nov. 2018.

MICHAEL CHINEN (Member, IEEE) received the B.S. degree in computer science from the University of Washington, the B.M. degree in music from the University of Washington, in 2005, with a focus on computer music and synthesis, and the M.A. degree in digital music from Dartmouth, in 2009. After his B.M. degree, he spent the following two years as a Research Student at Tokyo Denki University under Naotoshi Osaka's Sound Media Representation Laboratory, researching analysis-synthesis frameworks. After his M.A. degree, he was a Fulbright Scholar with the Technical University of Berlin immediately following. He currently works with Google, focusing on speech coding and quality.