SQAPP: NO-REFERENCE SPEECH QUALITY ASSESSMENT VIA PAIRWISE PREFERENCE

Pranay Manocha[1], *Zeyu Jin*[2], *Adam Finkelstein*[1]

[1]Princeton University, USA [2]Adobe Research, USA

ABSTRACT

Automatic speech quality assessment remains challenging, as we lack complete models of human auditory perception. Many existing *full-reference* models correlate well with human perception, but cannot be used in real-world scenarios where ground truth clean reference recordings are not available. On the other hand *no-reference* metrics typically suffer from several shortcomings, such as lack of robustness to unseen perturbations and reliance on (limited) labeled data for training. Moreover, noise or large variance among the labels makes it difficult to learn generalizable representations, especially for recordings with subtle differences. This paper proposes a learning framework for estimating the quality of a recording *without* any reference, and without any human judgments. The main component of this framework is a *pairwise quality-preference* strategy that reduces label noise, thereby making learning more robust. From pairwise preferences, we first learn a content invariant quality ordering; and then we re-target the model to predict quality on an absolute scale. We show that the resulting learned metric is well-calibrated with human judgments. Since it is a deep network, the metric is *differentiable*, making it suitable as a loss function for downstream tasks. For example, we show that adding this metric to an existing speech enhancement method yields significant improvement.

***Index Terms*—** audio quality, speech quality, no-reference metric, perceptual metric, pairwise preference, speech enhancement

1. INTRODUCTION

Speech quality assessment (SQA) plays a fundamental role in many applications affecting the quality of listening experiences. Many factors affect perceived speech quality, including audio codecs, network conditions, and background noise. Gold standard assessments of speech quality requires subjective listening tests. For instance, in the ITU standard P.800, participants judge speech quality on a 5-point scale. The sound quality of a stimulus is then measured with the *mean opinion score* (MOS) from several listeners. Such subjective evaluations are time consuming and costly, especially when repeated many times per recording, and are therefore not scalable. Thus automatic (objective) SQA methods are more practical.

Several *full-reference* objective methods have been developed: PESQ [1], POLQA [2], VISQOL [3], DPAM [4] and CDPAM [5] that produce a quality rating by comparing a corrupted speech signal to its clean reference. These methods correlate well with human judgment, but the requirement of a paired clean reference limits applicability in typical real-world scenarios. Moreover, since these *similarity* metrics estimate the distortion with respect to a reference, they can fail to reflect absolute *quality*. Fig 1 illustrates a scenario where two high quality reference recordings from different acoustic settings lead to different relative similarity (quality) measures for a third lower-quality recording. Thus, full-reference quality measures evaluated with respect to a particular clean reference set (e.g. DAPS [6]) may give different ratings than one using a different (also clean) reference set (e.g. VCTK [7]).

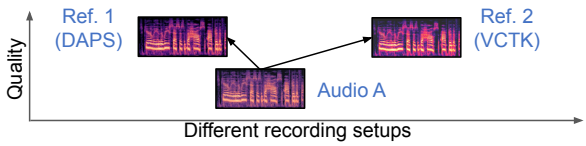

Fig. 1: **Similarity vs. Quality**: Ref.1 and Ref.2 are two recordings from different datasets (e.g. DAPS [6], VCTK [7]), of equal (high) quality but with different acoustic settings. Audio A, a lower quality recording from a different setting, may have different similarity (distance) to the two references, even though they have equal quality. Arrow length indicates similarity measure.

No-reference methods avoid these limitations by rating quality on an absolute scale. Traditional methods like ITU standard P.563 [8] and SRMRnorm [9] involve complicated hand-crafted features. State of the art approaches rely on deep learning [10–18]. Older learning methods trained models on objective scores (e.g. PESQ) [11], while more recent approaches discover a mapping between noisy audio signals and MOS in a supervised learning fashion [10, 16–18]. However, as observed by Manocha et al. [19], no-reference metrics learn an implicit distribution of clean references, which suffers from both (a) high variance due to factors like mood and past experience, and (b) substantial label variance from human annotations. For example, in DNSMOS [10], almost half of the recordings have ratings with standard deviation > 1. This label noise poses challenges in training robust models. Collecting a consistent MOS dataset faces many challenges including the need for uniform listening setups across many subjects; and even then ratings may not be consistent when repeated. Finally, MOS ratings depend on the conditions of the stimuli, which are task-dependent (making it difficult to combine MOS ratings across tasks).

Instead, our method learns directly from paired-preference data. The inspiration for this approach comes from the observation that humans find preference assessment easier than assigning ratings on an absolute scale [20]. Thus, preference ratings tend to be more robust and repeatable, and have lower variance [21]. Therefore, our hypothesis is that models trained on these relative assessments will likely lead to better generalization. Another benefit of the pairwise-preference approach is that it can completely rely on synthetically produced data – for example a recording with two different levels of noise added, where we *assume* that more noise means lower quality.

We propose SQAPP: No-reference Speech Quality Assessment via Pairwise Preference. We first train an audio *preference* predictor model that learns content-invariance; next, we adapt it to a *rating* predictor model that provides a rating of acoustic quality on an absolute scale. We evaluate SQAPP comprehensively by correlating it with 20 existing publicly available datasets: it compares favorably to existing baselines, even without relying on *any* subjective labels. Moreover, since SQAPP relies only on simulated data, it can be easily adapted to situations where there is a mismatch in train and test degradations. Finally, we show that adding SQAPP to the loss function yields significant improvements to the existing state of the art model for speech enhancement [22]. The code, resulting metric, as well as listening examples, are available here:
`https://pixl.cs.princeton.edu/pubs/Manocha_2022_SNS/`

2. THE SQAPP FRAMEWORK

Our framework is designed to assess the quality of a speech recording without any reference. We train the SQAPP metric in two stages, depicted in Fig 2: we first train a pairwise quality *preference prediction* model; and, second, we re-target the learned embeddings to a single-input *rating prediction* model. A higher rating signifies that the corresponding recording is high quality.

2.1. Architecture

SQAPP's architecture (Fig 2) comprises of four modules: an embedding block, a temporal-aggregation block, a preference predictor and a rating predictor. In the first stage, we train the embedding and temporal-aggregation blocks with a preference predictor that predicts which one of the two audio inputs have better quality. In the second stage we replace the preference predictor with a rating predictor to output a quality rating for a single audio input.

Embedding block: We use the Inception [23] architecture to extract a feature embedding. The 6-block Inception network consists of 64 conv. filters, followed by 1x1, 3x3 and 5x5 filters, leading to 3x3 max-pooling, and a 1x2 max-pool along the frequency dimension.

Temporal Aggregation: We use Temporal Convolutional Networks (TCNs) [24] in this block. While the phonetic speech content can change per frame, the acoustics (recording conditions, background noise, distortions, etc) require capturing long-term history for quality assessment. The network consists of 4 temporal blocks, with each block consisting of 2 convolutional layers with a kernel size of 1x3. Each layer consists of 32, 64, 64 and 128 channels for each of the 4 blocks respectively. Each layer uses weight normalization, and dilated convolutions with dilation factors of 2, 4, 8 and 16 for the 4 blocks.

Preference predictor: In stage 1, we construct a *Preference predictor* which takes the two temporally aggregated embeddings (output of the above two stages) as input, and outputs a logit at every frame that indicates which input recording of the two has a higher quality. This network consists of 3 convolutional layers, each consisting of 32, 8 and 2 channels respectively. The kernel size for each layer is 1x5, with each layer having BatchNorm and Dropout.

Rating predictor: In stage 2, we replace the *Preference predictor* with a *Rating predictor* that takes in a single temporally aggregated embedding and outputs a logit per frame, representing the absolute quality. This module consists of 4 convolutional layers, each consisting of 64, 32, 8 and 2 channels respectively. The kernel size for each layer is 1x5, with each layer having BatchNorm and Dropout.

2.2. Dataset Generation

To obtain pairwise preference data, we assume the availability of a clean speech database $\mathcal{D}_c$. Let $\mathbf{x}_{AB} = (x_A, x_B)$ be an ordered pair input to the network. Two clean recordings s_A and s_B are sampled from $\mathcal{D}_c$, and are corrupted to produce x_A and x_B. This is done by sampling a noise from a collection of noises, and adding it to s_A and s_B at two different levels uniformly sampled from a range (Table 1). The hard-target-label $\mathbf{y}_{AB}$ for $\mathbf{x}_{AB}$ is $[1, 0]$ if noise level of x_A is lower than x_B and $[0, 1]$ otherwise.

Note here that the inputs to be model ($\mathbf{x}_A$ and $\mathbf{x}_B$) are created by sampling two *different* clean recordings, but adding the *same* noise at two different levels. This encourages the preference predictor model at stage 1 to be invariant to recording content, which further empowers training of a robust rating predictor in stage 2. We also use PESQ to calibrate across different perturbations, and serve as a proxy for human annotations. These findings can be treated as a pilot study for future work on a dataset of subjective quality preferences.

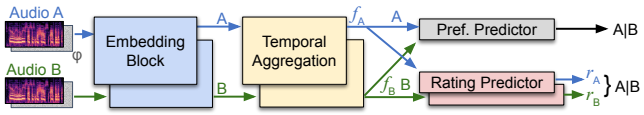

Fig. 2: **Training architecture**: (i) We first train a pairwise quality *preference* predictor (gray) using x_A and x_B as inputs; and then (ii) re-target it to a single rating (*rating predictor* - pink) to output quality ratings r_A and r_B.

2.3. Training procedure

Stage 1: In each step, pairs of audio $\mathbf{x}_{AB} = (x_A, x_B)$ are passed through the embedding and aggregation blocks independently to obtain aggregated framewise embeddings f_A and f_B. The preference predictor network takes these two vectors as input and outputs two logits per frame to match $\mathbf{y}_{AB}$. We use a label smoothed version of cross-entropy loss [25], which encourages small logit gaps and prevents model over-fitting and overconfident predictions. If y_k and p_k denote the target and prediction respectively (y_k = 1 for the correct class, 0 otherwise), and α denotes the smoothing parameter, the loss is given by:

$$H(\mathbf{y}, \mathbf{p}) = \sum_{k=1}^{K=2} -(y_k(1-\alpha) + \frac{\alpha}{K}) \log(p_k) \quad (1)$$

Stage 2: We now detach the preference predictor network and pass f_A and f_B independently into the rating predictor network that outputs scalars r_A and r_B for each sample respectively. These two scalars are then rescaled with softmax before the smoothed cross-entropy loss is used to train the whole network. Because now all the modules in this network only use one audio as input, the output is a no-reference score reflective of absolute audio quality.

3. EXPERIMENTAL SETUP

3.1. Datasets and Training

For training and validation, we choose the clean speech recordings ($\mathcal{D}_c$) from the DNS Challenge [26] and DAPS [6]. FSDK50 [27] and the DNS Challenge Noise set serves as the noise dataset for additive noise degradations. Apart from additive noises, we also employ other possible degradations including hum noise, clipping, sound effects, packet losses, phase distortions, and a number of audio codecs. More details can be found in Table 1.

The inputs to the network are 3-second audio excerpts, represented by their Short-Time Fourier Transform (STFT). We stack together the magnitude and phase of the STFT as two channels. More specifically, STFTs are computed using a hamming window of size 32 ms with 50% overlap (at 16kHz sampling rate). Only the 256 positive frequencies are used. The label-smoothing parameter α (from Eq. 1)

Category	Perturbations	Intervals/Range
Additive [26, 27]	DNS-Challenge Noise Set FSDK50	-10 dB to 60 dB SNR
Reverb [26]	Direct to Reverberation Ratio (DRR) Reverberation Time (RT60)	-27 dB to 65 dB 0.05 sec to 8 sec
Compression	MP3 (bitrate) μ law compression (re-quantization)	8 Kb/sec to 320 Kb/sec 1 bit to 60 bits
Equalization	Frequency bands (cut/boost bands)	0 to 1
Miscellaneous	Pops (% audio samples) Griffin-lim (iterations) Buffering and Packet-loss Frequency Masking Resampling SoX-based (high-pass,lowpass,gain...) Clipping Dropouts (% audio samples)	0.01 % to 10 % 1 to 500 0.05 to 0.50 sec 0 to 0.50 4KHz to 16KHz 0 to 40 0.01 % to 20 %

Table 1: All examined perturbations and configurations.

Type	Name	HiFi [28]	VoCo [29]	FFTnet [30]	Noizeus [31]	PEASS [32]	Dereverb [33]	TCD-VoIP [34]	SASSEC [35]	SiSEC08 [36]	SAOC [37]	PEASS_db [35]
Full-ref.	PESQ	0.70	0.43	0.49	0.42	0.71	0.85	0.90	0.85	0.97	**0.76**	**0.42**
	CDPAM	0.68	**0.73**	0.68	**0.71**	**0.74**	**0.93**	0.88	**0.90**	**0.94**	0.63	0.35
Non-match.-ref	NORESQA	0.71	0.41	0.51	0.39	0.40	0.75	0.46	-	-	-	-
No-ref.	NISQA	**0.90**	0.29	0.48	0.63	0.29	0.81	**0.94**	0.63	0.66	0.62	0.25
	DNSMOS	0.88	0.48	0.53	0.59	0.21	0.73	0.72	0.60	0.40	0.32	-0.21
	SQAPP	0.88	0.49	**0.80**	0.45	0.39	0.86	0.90	0.79	0.69	0.54	0.38

Table 2: **MOS Correlations (1)**: Spearman correlations (SC) between various methods and MOS ratings are shown. ↑ is better.

is 0.25. We use Adam optimizer with a learning rate of 10^{-4} with an effective batch size of 128. We train the network for 1000 epochs using 4 Tesla V100 gpus. The initial weights of the model are chosen from the normal distribution $\mathcal{N}(0, 1e^{-2})$.

3.2. Baselines

We compare SQAPP with full-reference metrics like PESQ [1] and CDPAM [5]. CDPAM is a full-reference neural network based metric trained on a dataset of human annotated just-noticeable-difference (JND) ratings. Being full-reference, these methods require a clean reference for quality assessment. We also use DNSMOS [10] and NISQA [18] as no-reference SQA baselines. Both are trained on a large scale dataset of MOS ratings, however DNSMOS is focused on noise suppresors and NISQA on distortions that occur in communication networks. We also use NORESQA [19] as a baseline for assessing performance in a non-matched setting. It was also trained using simulated data, but using a different formulation that estimates the relative SNR and SI-SDR difference between two non-matched recordings. Also note that all metrics are evaluated at 16KHz except NISQA which is evaluated at 48KHz.

4. EXPERIMENTS

4.1. Evaluation

MOS Correlations: We use previously published diverse third-party studies to verify that our trained metric correlates well with their task. We evaluate our metric by computing Spearman's Rank Order Correlation (SC) of our metric's scores with MOS ratings on each dataset. These correlation scores are evaluated per speaker where we average scores for each speaker for each condition. In addition to all evaluation datasets considered by Manocha et al. [5, 19], we consider several additional datasets:

1. **SEBASS** [35–37] Subjective Evaluation of Blind Audio Source Separation: consists of five listening tests on audio quality of separated audio sources from various blind source separation systems. These listening tests are referred to as: **SASSEC**, **SiSEC08**, **PEASS_db**, **SiSEC18**, and **SAOC**. In each listening test, the listeners rated separated signals submitted as part of community-based signal separation evaluation.
2. **VCC16** [38]: consists of a large scale subjective evaluation of voice conversion (VC) systems across 17 teams who participated, with each team submitting 25 voice converted samples. Subjective evaluations were done to assess naturalness of the converted voice, as well as similarity between the converted voice to the target speaker. Here we only correlate with *naturalness*.
3. **HiFiGAN2** [40]: consists of subjective quality tests to assess audio quality across various denoising and dereverberation models including HiFi-GAN2, HiFi-GAN, FullSubNet and others. Subjective evaluations were conducted using the DAPS dataset (**H_Daps**) and using real-world recordings (**H_Ted**).

2AFC Tests: is a comparative approach to subjective evaluations. In this case, listeners are given a reference and two test recordings and asked to judge which one sounds more similar (in terms of quality) to the reference. We follow the evaluation protocol of Manocha et al. [5] and report how accurately different methods can predict same judgments as humans.

Results for the correlations and accuracy with subjective ratings are displayed in Tables 2, 3 and 4. First, we note that SQAPP is not only competitive to existing no-reference (DNSMOS and NISQA) and partial reference (NORESQA) baselines but even surpasses their performance in several cases (VoCo, FFTnet, PEASS, Dereverb, SASSEC, SiSEC08, PEASS_db, VCC16, VCC18 and HiFi2). Note that both DNSMOS and NISQA are explictly trained on large scale MOS datasets, whereas our method is not trained on any perceptual judgments. Second, similar to findings by Manocha et al. [4, 5], conventional metrics like PESQ perform better on measuring large distances (e.g. *Dereverb*) than subtle differences (e.g. phase: *FFTnet*). This is in contrast to SQAPP that performs coequally for both large and small differences. Third, we note that NISQA scores better correlations than DNSMOS across various datasets. As described earlier, large label noise poses a big challenge in training robust models. The DNSMOS subjective dataset is composed of ratings from at-most 10 annotators, whereas NISQA has at-most 30 annotators per recording. This ensures that the latter has lower label noise, and therefore is a more robust model. This is also precisely the motivation behind our proposed quality-*preference* framework. Fourth, we see that the performance gap between SQAPP and full-reference methods (like PESQ and CDPAM) is small, suggesting that SQAPP can be a good substitute. Also since SQAPP is a no-reference metric, it stays useful in practical and real-world situations where the clean reference might not be available. Moreover, our approach is especially useful for datasets where the test degradations are different from train degradations. For e.g., both NISQA and DNSMOS show low correlations for PEASS,

Type	Name	VCC16 [38]	VCC18 [39]	H_Ted [40]	H_Daps [40]	SiSEC18 [35]
Full-ref.	PESQ	0.37	0.22	0.88	0.67	**0.90**
	CDPAM	0.45	0.63	**0.94**	0.54	0.68
Non-match-ref.	NORESQA	-	0.41	-	-	-
No-ref.	NISQA	0.39	0.32	**0.94**	0.72	0.08
	DNSMOS	-0.21	0.42	0.89	0.69	0.40
	SQAPP	**0.61**	**0.64**	0.89	**0.97**	0.31

Table 3: **MOS Correlations (2)**: Spearman correlations (SC) between various methods and MOS ratings are shown. ↑ is better.

Type	Name	Simulated	FFTnet	BWE	HiFi
Full-ref.	PESQ	86.0	67.0	38.0	88.5
	CDPAM	**87.7**	88.5	75.9	86.5
Non-match-ref.	NORESQA	68.7	73.3	53.3	81.6
No-ref.	NISQA	79.8	**91.8**	61.1	**94.4**
	DNSMOS	49.2	58.8	45.0	62.3
	SQAPP	84.5	81.1	**78.1**	91.5

Table 4: **2AFC accuracy** showing % accuracy of various methods with subjective ratings. ↑ is better.

Name	HiFi	Dereverb	PEASS	VCC16	H_Ted
SQAPP- *default*	0.88	**0.86**	**0.39**	**0.61**	0.89
after stage 1	**0.93**	0.75	0.33	0.28	**0.95**
w/o PESQ calib.	0.85	0.73	0.18	0.44	0.77
same clean rec.	0.40	0.68	0.35	0.21	0.79
w/o phase rep.	0.49	0.77	0.19	0.56	0.88
target PESQ	0.20	0.59	-0.21	-0.12	0.14
w/o stage 1	0.55	0.65	0.08	0.10	0.77

Table 5: **Ablation studies**. Spearman correlations (SC) of our model with MOS across different ablation components are shown (Sec 4.2). ↑ is better.

VCC16 and VCC18, whereas our metric obtains better correlations. Fifth, SQAPP performs quite favorably to NORESQA which is also trained entirely using simulated data. This is because predicting objective metrics (e.g. SNR and SI-SDR) is non-trivial, especially for cases where different recordings with different perturbations are compared. Finally, looking beyond MOS Correlations, SQAPP outperforms the baselines (DNSMOS and NORESQA) by a considerable margin. This shows that SQAPP not only generalizes well to MOS ratings, but also does well on different perceptual tests beyond MOS (e.g. 2AFC). It's competitive with full-reference baselines and NISQA which is designed to work at 48KHz.

4.2. Ablations

We perform ablation studies to better understand the influence of different components of our metric to address our various design choices (Table 5). We first compare our final model (*default*) with models after stage 1 (*preference model*). We also contrast the results of our final model with other models that: (i) do not use PESQ calibration; (ii) use a different feature representation as input to the model; and (iii) are trained using different objectives and formulation. These models are compared on correlation with subjective ratings from a subset of existing datasets.

Different components of SQAPP: Each of the two stages of SQAPP namely pairwise *preference* (*after stage-1 full-ref.*) and *rating prediction* models (*default*) show good correlations with subjective ratings. Even the pairwise preference model acting as full-reference shows good correlation with subjective ratings. Both our full-reference and no-reference models perform favorably to existing baselines, which suggests the usefulness of our framework, even without any training on perceptual data.

No PESQ Calibration:(*w/o* PESQ *calib.*) For models that do not use PESQ calibration, we observe a significant drop in correlations across datasets including HiFi, VCC16, VCC18 and HiFi2. These are typically datasets that compare many methods, with each method having a different artifact. This is a result of our training strategy where the ordered input pair has the same noise, but at two different levels, since it allows for a more clear comparison of quality. All other datasets show non-significant change in correlations. This shows that PESQ helps in calibrating the ratings across perturbations by acting as a proxy for human judgments.

Different input feature representation: To evaluate how the performance of the metric varies with different input features, we train two sets of models: (i) with the exact *same* clean recording for training (*same clean rec.*); and (ii) without using phase features as input to the model (*w/o phase rep.*). For (i), we observe that the correlations across downstream datasets are usually lower suggesting that enforcing the content in-variance property helps learn robust acoustic quality features. For (ii), we observe that several datasets (e.g. HiFi) that utilize the phase information to predict quality show low correlation with subjective ratings. This shows that our *default* model makes use of phase features to predict quality.

Different objectives and formulation: We also investigate the choice of objectives and formulation, where we look at another formulation of a no-reference metric by directly training on PESQ [11] (*target* PESQ). However, this approach suffers from mode-collapse and variance [41] in the output distribution, wherein the model outputs the mean of the PESQ score for a test recording that has more than one possible references (see Fig1). The performance of this model is lower than our proposed approach. Additionally, we also investigate the importance of the pairwise-preference task (*w/o stage 1*), wherein we train a model *directly* to predict a rating without first enforcing content in-variance. The correlations with subjective ratings are lower, and our hypothesis is that the metric may generalize poorly to unseen speakers or content [5]. This also suggests the usefulness of the *pairwise-preference* model of our approach.

	PESQ	STOI	CSIG	CBAK	COVL	MOS
Noisy	1.97	91.50	3.35	2.44	2.63	2.722±0.014
DEMUCS [22]	3.01	95.00	4.39	3.44	3.73	4.075±0.015
CDPAM [5]	3.06	94.93	4.30	**3.56**	3.70	4.082±0.013
SQAPP	**3.09**	**94.99**	**4.40**	3.49	3.78	**4.339**±0.018

Table 6: **Evaluation of denoising** models using the VCTK [7] test set with five objective measures and MOS score. As reference, the clean samples have a MOS score of **4.511**±0.011

4.3. Speech Enhancement

To further demonstrate the effectiveness of our metric, we use the current state-of-the-art DEMUCS architecture based speech denoiser [22] and supplement SQAPP metric by adding it as an additional loss while training. The training dataset consists of VCTK [7] and DNS [42] datasets.

For the evaluation, we randomly select 500 recordings from the VCTK test set, and then perform speech enhancement. We then conduct MOS test on enhanced samples using Amazon Mechanical Turk (AMT). For the objective assessments, we use: i) PESQ (from 0.5 to 4.5); (ii) Short-Time Objective Intelligibility (*STOI*) (from 0 to 100); (iii) *CSIG*: MOS prediction of the signal distortion attending only to the speech signal (from 1 to 5); (iv) *CBAK*: MOS prediction of the intrusiveness of background noise (from 1 to 5); (v) *COVL*: MOS prediction of the overall effect (from 1 to 5). We compare our model with the baseline approach, as well as using CDPAM [5] as a loss function. Results are shown in Table 6.

For subjective studies, we conducted a MOS listening study on AMT where each subject is asked to rate the sound quality of an audio snippet on a scale of 1 to 5, with *1=Bad, 5=Excellent*. In total, we collect around 3240 ratings for each method over 159 unique workers. We provide studio-quality audio as reference for high-quality, and the input noisy audio as low-anchor. We observe that our SQAPP-trained enhancement model scores the highest MOS score amongst all baselines. Specifically, SQAPP can identify and eliminate minor human perceptible artifacts that are not captured by traditional losses. Moreover, it also trains faster as compared to CDPAM(16 hours as compared to 19 hours). This highlights the usefulness of using SQAPP in audio quality tasks.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented SQAPP - a framework for no-reference speech quality assessment. We showed the utility of our pairwise *preference* strategy that reduces label noise present in human annotations, as well as enforces content in-variance that helps learn a robust absolute rating model. In subjective evaluations, our method works as competently as established full-reference, partial reference and no-reference SQA methods. At the same time, it addresses key limitations of those methods. Going forward, our focus will be on developing novel methods within this framework which can correlate better with subjective human ratings.

6. REFERENCES

[1] A. W. Rix, J. G. Beerends, et al., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, vol. 2, pp. 749–752.

[2] J. G. Beerends, C. Schmidmer, et al., "Perceptual objective listening quality assessment (POLQA), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of the AES*, , no. 6, 2013.

[3] A. Hines, J. Skoglund, et al., "ViSQOL: An objective speech quality model," *EURASIP*, vol. 2015, no. 1, 2015.

[4] P. Manocha, A. Finkelstein, et al., "A differentiable perceptual audio metric learned from just noticeable differences," *Interspeech*, 2020.

[5] P. Manocha, Z. Jin, et al., "CDPAM: Contrastive learning for perceptual audio similarity," ICASSP 2021.

[6] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE SPS*, vol. 22, no. 8, 2014.

[7] C. Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and TTS models," 2017.

[8] L. Malfait, J. Berger, et al., "P. 563—the ITU-T standard for single-ended speech quality assessment," *IEEE TASLP*, vol. 14, no. 6, pp. 1924–1934, 2006.

[9] J. F. Santos, M. Senoussaoui, et al., "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *IEEE IWAENC*, 2014, pp. 55–59.

[10] C. K. Reddy, V. Gopal, et al., "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *ICASSP*, 2020.

[11] S.-W. Fu, Y. Tsao, et al., "Quality-Net: end-to-end non-intrusive speech quality assessment model on blstm," *Interspeech*, 2018.

[12] A. H. Andersen, J. M. De Haan, et al., "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM TASLP*, vol. 26, no. 10, pp. 1925–1939, 2018.

[13] H. Gamper, C. K. Reddy, et al., "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *WASPAA*, 2019, pp. 85–89.

[14] A. R. Avila, H. Gamper, et al., "Non-intrusive speech quality assessment using neural networks," in *ICASSP*, 2019.

[15] M. Yu, C. Zhang, et al., "MetricNet: Improved modeling for non-intrusive speech quality assessment," *Interspeech*, 2021.

[16] Z. Zhang, P. Vyas, et al., "An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework," in *ICASSP*, 2021, pp. 316–320.

[17] A. A. Catellier and S. D. Voran, "WaweNets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *ICASSP*, 2020.

[18] G. Mittag, B. Naderi, et al., "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Interspeech*, 2021.

[19] P. Manocha, B. Xu, et al., "NORESQA: A framework for speech quality assessment using non-matching references," *NeurIPS 2021*, vol. 34.

[20] L. L. Thurstone, "A law of comparative judgment.," *Psychological review*, 1927.

[21] A. R. Teodorescu, R. Moran, et al., "Absolutely relative or relatively absolute: violations of value invariance in human decision making," *Psychonomic bulletin*, 2016.

[22] A. Defossez, G. Synnaeve, et al., "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[23] C. Szegedy, W. Liu, et al., "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[24] S. Bai, J. Z. Kolter, et al., "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[25] R. Müller, S. Kornblith, et al., "When does label smoothing help?," in *NeurIPS*, 2019.

[26] C. K. Reddy, H. Dubey, et al., "Interspeech 2021 deep noise suppression challenge," in *INTERSPEECH*, 2021.

[27] E. Fonseca, X. Favory, et al., "FSD50k: an open dataset of human-labeled sound events," *arXiv preprint*, 2020.

[28] J. Su, Z. Jin, et al., "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *Interspeech*, 2020.

[29] Z. Jin, G. J. Mysore, et al., "VoCo: Text-based insertion and replacement in audio narration," *ACM TOG*, 2017.

[30] Z. Jin, A. Finkelstein, et al., "FFTNet: A real-time speaker-dependent neural vocoder," in *ICASSP*, 2018.

[31] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech comm.*, 2007.

[32] V. Emiya, E. Vincent, et al., "Subjective and objective quality assessment of audio source separation," *IEEE TASLP*, 2011.

[33] J. Su, A. Finkelstein, et al., "Perceptually-motivated environment-specific speech enhancement," in *ICASSP*, 2019.

[34] N. Harte, E. Gillen, et al., "TCD-VoIP, a research database of degraded speech for assessing quality in voip applications," in *QoMEX*, 2015.

[35] T. Kastner and J. Herre, "An efficient model for estimating subjective quality of separated audio source signals," in *WASPAA*, 2019, pp. 95–99.

[36] T. Kastner, "Evaluating physical measures for predicting the perceived quality of blindly separated audio source signals," in *AES Convention*, 2009.

[37] J. Breebaart, J. Engdegård, et al., "Spatial audio object coding (SAOC)-the upcoming mpeg standard on parametric object based audio coding," in *AES Convention 124*. AES, 2008.

[38] T. Toda, L.-H. Chen, et al., "The Voice Conversion Challenge 2016," in *Interspeech*, 2016, pp. 1632–1636.

[39] J. Lorenzo-Trueba, J. Yamagishi, et al., "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *arXiv:1804.04262*, 2018.

[40] J. Su, Z. Jin, et al., "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," WASPAA, 2021.

[41] X. Dong and D. S. Williamson, "A classification-aided framework for non-intrusive speech quality assessment," in *WASPAA*, 2019.

[42] C. K. Reddy, E. Beyrami, et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," in *Interspeech 2020*.