A CLASSIFICATION-AIDED FRAMEWORK FOR NON-INTRUSIVE SPEECH QUALITY ASSESSMENT

Xuan Dong[‡] and Donald S. Williamson[‡]

[‡]Department of Computer Science, Indiana University - Bloomington, USA

Problem Overview

Intrusive objective metrics, such as the perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), have become standard measures for evaluating speech. These metrics enable efficient and costless evaluations, where ratings are often computed by comparing a degraded speech signal to its underlying clean reference signal. However, they cannot be used to evaluate real-world signals that have inaccessible references.

Non-intrusive objective metrics perform evaluations directly on the signal of interest, without the need for a reference signal. These metrics rely on properties of signals or environmental factors to determine quality and intelligibility scores. Current non-intrusive metrics have many limitations, including:

- they perform worse than intrusive measures in terms of correlations to human listening evaluations
- they have not been thoroughly evaluated in realistic environments that contain many speakers or different types of acoustical noise
- they are only intended for specific-signal types
- their prediction are not reliable in very low SNR conditions since the estimation error and variance are high

Motivation

Related works

Data-driven approaches have been proposed recently as a means of evaluating speech quality, intelligibility, naturalness, and mean opinion score:

- machine learning techniques: classification and regression trees [Sharma *et al.* 2016]
- deep learning approaches: deep neural network [Ooster *et al.* 2018], convolutional neural network [Andersen *et al.* 2018], a stack of long short-term memory [Patton *et al.* 2016], bidirectional long short-term memory [Fu *et al.* 2018]

These approaches are promising since they enable quick reference-less evaluations, and the algorithms learn from data without prior assumptions.

Our idea

Inspired by the latter deep-learning based metrics, we propose a convolutional neural network (CNN) framework for assessing the perceptual quality of speech. More specifically, we jointly train a CNN to predict the categorical objective ranking and true PESQ score, where PESQ scores are grouped into categorical classes based on pre-defined ranges.

Hence, we propose to treat objective speech evaluation as the combination of a classification and a regression task. The two tasks share the same feature extraction layers while each task also has independent modules to achieve specific goals. Learning tasks in parallel while using a shared representation has been shown to be helpful for other multi-task learning problems.

Model

Network architecture: our utterance-level classification-aided nonintrusive (**UCAN**) assessment approach uses a multi-layered CNN to predict both the categorical quality rankings of noisy speech and the corresponding objective quality scores.

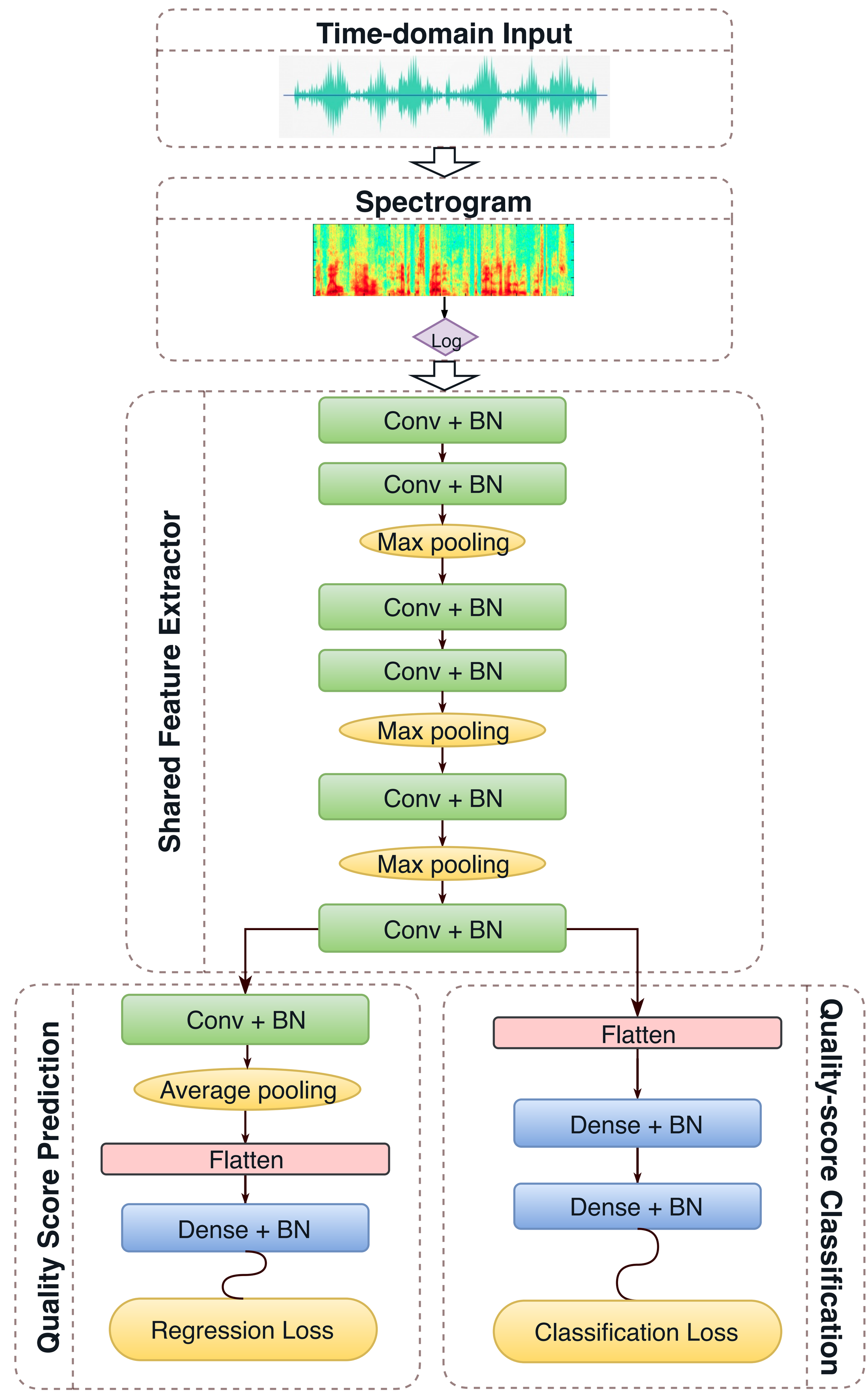

Fig. 1: Architecture of the proposed framework with shared convolutional and task-specific fully connected layers.

PESQ quality labels: Two training targets are simultaneously ~~applied~~ in our model. One is the raw PESQ score $\mathcal{S}_{pesq}$ for a particular signal, and the other is the corresponding ~~the~~ quality class. The PESQ classification label of a given signal is calculated by

$$\text{Class}(\mathcal{S}_{pesq}) = \min(\max\left(1, \text{ceil}\left(\frac{\mathcal{S}_{pesq} - L_t}{B}\right)\right), N), \tag{1}$$

where L_t denotes the low threshold, B denotes the category bin size~~, and ceil(·) denotes the ceiling function.~~

Objective function: the mean squared loss (regression loss $\mathcal{L}_{regr}$) that stems from the left subnet together with the cross entropy loss (classification loss $\mathcal{L}_{cls}$) are utilized to update the weights of the shared network:

$$\mathcal{L}_{total} = \beta * \mathcal{L}_{cls} + (1 - \beta) * \mathcal{L}_{regr}, \tag{2}$$

where β controls the trade-off between optimizing the network for the classification or regression task.

Comparison

- We used 25,000 training mixtures, and 15,000 testing mixtures that are generated from TIMIT corpus and NOISEX-92 noise database
- Cover a wide range of SNRs: from -25 dB to 30 dB with 5 dB increments

	Seen noisy speech		Unseen noisy speech		Enhanced speec	
	MSE	PCC	MSE	PCC	MSE	PCC
NISA [Sharma *et al.* 2016]	0.156	0.86	0.183	0.84	0.151	0.88
DESQ [Ooster *et al.* 2018]	0.170	**0.91**	0.246	**0.90**	0.168	0.91
CNN [Andersen *et al.* 2018]	**0.139**	0.89	0.185	0.86	0.123	0.90
AutoMOS [Patton *et al.* 2016]	0.162	0.88	0.391	0.85	0.175	0.90
Quality-Net [Fu *et al.* 2018]	0.149	0.90	**0.170**	0.89	**0.102**	**0.93**
UCAN ($\beta = 0$)	0.097	0.94	0.112	0.92	0.087	0.94
UCAN ($\beta = 0.2$)	**0.078**	**0.95**	**0.096**	**0.93**	**0.062**	**0.96**

Tab. 1: Performance comparison on seen and unseen conditions.

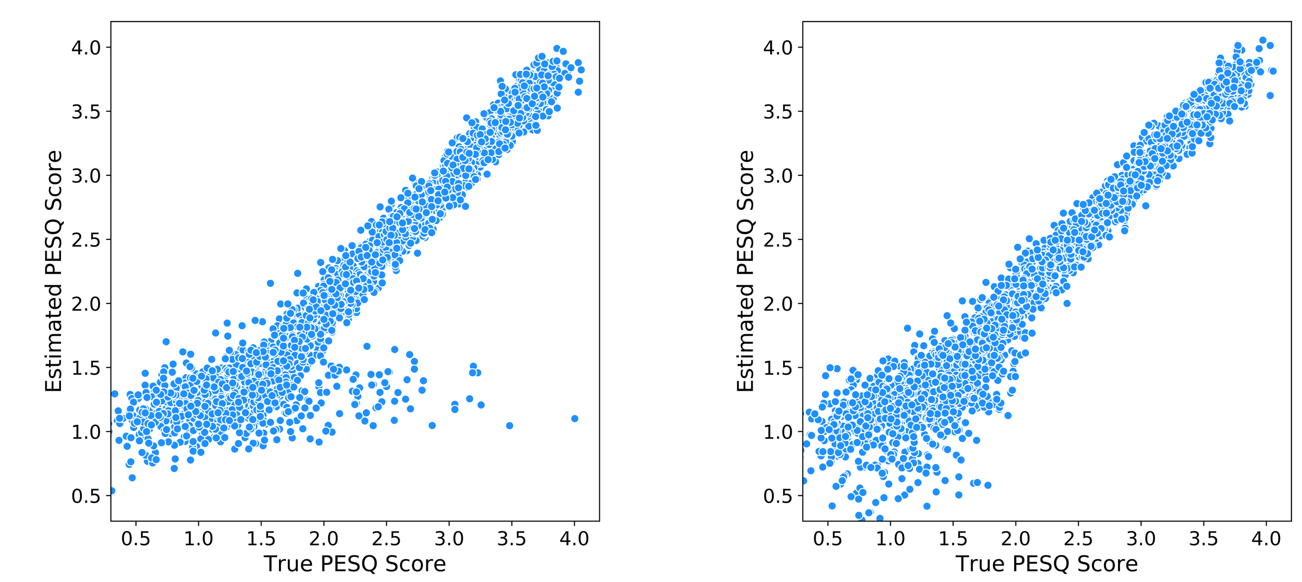

Fig. 2: Scatter plots of the true and the estimated PESQ scores on seen noise condition. From left to right: UCAN without ($\beta = 0$) or with ($\beta = 0.2$) classification-aided module.

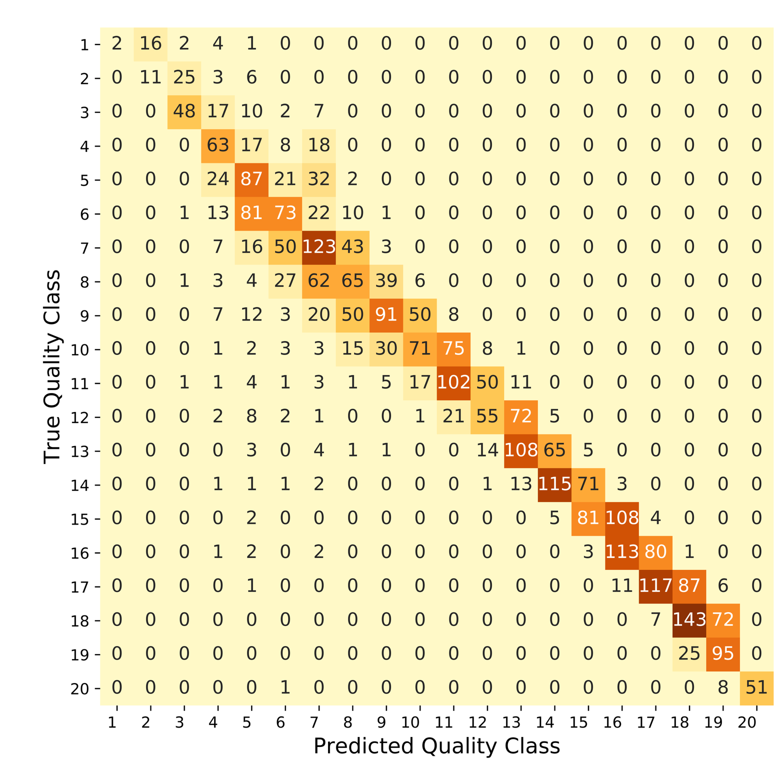

Fig. 3: Confusion matrix of the categorical classification task.

Conclusion

We present an utterance-level classification-aided non-intrusive speech quality assessment approach to predict both the objective quality class and the quality score of noisy and enhanced speech signals. The performance of UCAN outperforms previous state-of-the-art approaches, and significantly lowers estimation errors, which indicates that jointly training a classification-aided regression module is promising for speech quality assessment.