More for Less: Non-Intrusive Speech Quality Assessment with Limited Annotations

Alessandro Ragano[1,2], Emmanouil Benetos[3,4], and Andrew Hines[1,2]
[1] *School of Computer Science, University College Dublin, Ireland* [2] *Insight Centre for Data Analytics, Ireland*
[3] *School of EECS, Queen Mary University of London, UK* [4] *The Alan Turing Institute, UK*
alessandro.ragano@ucdconnect.ie, emmanouil.benetos@qmul.ac.uk, andrew.hines@ucd.ie

***Abstract*—Non-intrusive speech quality assessment is a crucial operation in multimedia applications. The scarcity of annotated data and the lack of a reference signal represent some of the main challenges for designing efficient quality assessment metrics. In this paper, we propose two multi-task models to tackle the problems above. In the first model, we first learn a feature representation with a degradation classifier on a large dataset. Then we perform MOS prediction and degradation classification simultaneously on a small dataset annotated with MOS. In the second approach, the initial stage consists of learning features with a deep clustering-based unsupervised feature representation on the large dataset. Next, we perform MOS prediction and cluster label classification simultaneously on a small dataset. We show that the deep clustering-based model outperforms the degradation classifier-based model and the 3 baselines (autoencoder features, P.563, and SRMRnorm) on TCD-VoIP and P.Sup23 Exp1. In particular, the deep clustering-based approach shows good domain adaptation performance on P.Sup23 Exp1 which consists of degradations different from those included in the large dataset. This paper shows that multi-task learning combined with feature representations from unlabelled data is a promising approach to deal with the lack of large MOS annotated datasets.**

***Index Terms*—non-intrusive speech quality, multi-task learning, unsupervised feature representation, deep clustering**

I. Introduction

Speech quality assessment is fundamental to improve users' quality of experience (QoE) of multimedia communication systems. Perceived audio quality is affected by several degradations caused by many factors including audio codecs, network conditions, speech enhancement and background noise. The most accurate way to assess audio quality is through subjective listening tests. For instance, in the ITU standard P.800 [1], participants judge audio quality on a 5-point scale. Next, the sound quality of a stimulus is measured with the mean opinion scores (MOS) computed from several listeners.

Despite their reliability, subjective listening tests are not always convenient given that they require a (1) substantial number of participants; (2) they cannot be used in real-time applications; (3) they are not suitable when large sound collections have to be evaluated; (4) they can be time-consuming and expensive.

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 17/RC-PhD/3483 and 17/RC/2289P2 and was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. The work of EB was supported by RAEng Research Fellowship RF/128 and a Turing Fellowship.

Objective quality metrics are a reliable replacement of the subjective judgement in the conditions above. Objective quality metrics can be divided into *full-reference metrics*, where a reference signal is available, and *non-intrusive metrics* where quality is estimated through the noisy signal only. Non-intrusive objective metrics are preferred in scenarios where the reference signal does not exist such as real-time applications and real-world recordings. Traditional non-intrusive methods include the ITU standard P.563 [2] and SRMRnorm [3]. A more recent approach to predict audio quality is to learn a mapping between noisy audio signals and MOS in a supervised learning fashion. The main drawback of using supervised learning is that a considerable amount of annotated data is required. Annotating large datasets is a general problem in machine learning, especially in multimedia quality assessment where multiple recruiters are needed to annotate only one stimulus. Although annotations obtained through crowdsourcing can be as valid as a lab setting quality measures [4], [5], annotating data is still a costly and time-consuming operation.

To overcome the scarcity of annotated data, we propose unsupervised feature learning combined with multi-task learning. We introduce two multi-task learning techniques. In the first approach, we optimize both degradation classification and quality prediction simultaneously. Given the lack of large MOS annotated datasets, we propose to initialize the multi-task model by using the weights learned from a degradation classifier trained on a large dataset. Unlike quality prediction, classifying degradations can be done using large synthetic datasets where we apply various degradations. In the second approach, we propose a semi-supervised multi-task feature learning model without using the degradation labels. We first learn an unsupervised feature representation using a deep clustering technique [6] on a large dataset where the MOS annotations are not given. Next, we cluster a small MOS annotated dataset using the feature representation and we use the cluster assignments as labels for the multi-task learning step. Our proposed approach can be especially useful in real-world scenarios where the knowledge of the degradation is not given and a large amount of real-world recordings is available [7], [8].

In this paper we make the following contributions;

- We propose multi-task learning for non-intrusive speech quality prediction using degradation classification as an auxiliary task.

- We propose a semi-supervised multi-task feature learning model where the labels for the auxiliary task are not given and are being generated using an unsupervised feature representation based on deep clustering called deep convolutional embedded clustering (DCEC) [6], [9].
- We show that deep clustering-based feature representation combined with multi-task learning achieves promising performance in non-intrusive speech quality assessment using small datasets annotated with MOS.

II. Related Work and Motivation

Recently, some non-intrusive metrics with deep learning techniques emerged. Only a few studies used large datasets annotated with MOS [10], [11] while others relied on annotations created with full-reference metrics [12]–[14] or hybrid annotations [15]. Another group of non-intrusive metrics is closer to our approach, in the sense that they rely on different tasks to improve quality prediction. Ooster et al. [16] used an automatic speech recogniser, assuming that phoneme posterior probabilities from a neural network degrade in presence of factors that affect speech quality. Semi-Supervised Speech Quality Assessment (SESQA) [17] uses 5 complementary auxiliary tasks and 3 optimization criteria (MOS error, pairwise ranking, and score consistency). Soni et al. [18] use a fully connected autoencoder to learn a feature representation from a large dataset. To the best of the authors' knowledge, the study of Soni et al. is the only one that learns an unsupervised feature representation for speech quality prediction.

Multi-task learning [19] is based on training multiple tasks simultaneously. The motivation is that sharing weights between related tasks can improve all the tasks together. In our setting, MOS prediction is the main task and degradation classification or cluster assignment prediction is the auxiliary task. Multi-task learning improves generalisation and predictions of both tasks if the auxiliary task is related to the main task. Classifying degradations is a suitable auxiliary task for speech quality prediction because; (1) Perceived audio quality depends on the degradation [20]. The model has to learn how a clean speech signal is degraded, which is a concept associated with quality as well; (2) Classifying degradations is related to quality prediction but these tasks are not identical, which is desired [21]. Indeed, two completely different degradations might be annotated with the same MOS. Also, using degradation information together with quality prediction has been proposed in image quality assessment [22] and can be transferred to speech quality assessment similarly. In the second proposed approach, instead of relying on degradation labels, we generate cluster labels from unsupervised feature representations. The perceived quality of a speech signal is not only related to the degradation and more factors are involved. We assume that having labels that represent a semantic similarity between the data points could be more meaningful than using only degradation labels in a multi-task learning scenario. These annotations are generated using only unlabelled data through a deep clustering [6] technique that here we propose as a feature learning step for speech quality prediction. It must be noted that different clustering techniques for unsupervised learning of features have been already employed in computer vision [23], [24]

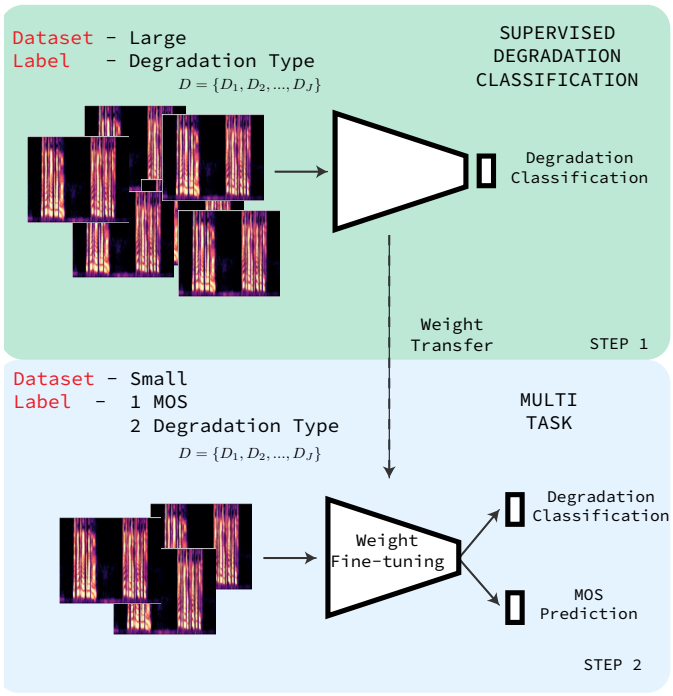


Fig. 1. MTL-based model. STEP 1 consists of training a supervised degradation classifier on a large dataset. STEP 2 is a multi-task network trained on a small MOS annotated dataset that simultaneously classifies degradations and predicts quality scores. The model in STEP 2 is initialized with the weights learned in STEP 1.

III. Method

In this section, we describe two methods for non-intrusive speech quality assessment based on Multi-Task Learning (MTL) and Semi-Supervised Multi-Task Learning (SEMTL).

A. Multi-Task Learning

In the MTL-based model, we train a model that performs simultaneous learning of MOS scores and degradation type. To tackle the lack of large MOS annotated datasets, we use the weights learned by a degradation classifier that is trained on a different and larger dataset. The approach is shown in Fig. 1. In the first step, we minimize the cross-entropy $\mathcal{L}_{ce}(D, \hat{D})$, where D represents the degradation class and $\hat{D}$ is the predicted degradation type. We use a large dataset to learn a rich feature representation so that we can reuse the weights. In the second step, we initialize the weights from step 1 and we minimize the multi-task loss

$$\mathcal{L}_{tot} = \mathcal{L}_{ce}(D, \hat{D}) + \mathcal{L}_{mse}(S, \hat{S}) \quad (1)$$

where S is the annotated MOS, $\hat{S}$ is the predicted score, and $\mathcal{L}_{mse}$ is the mean squared error. This second step is carried out on the small dataset where MOS annotations are available.

B. Semi-Supervised Multi-Task Learning

Motivation: In the SEMTL-based model, we study whether a multi-task approach can be designed without using human-annotated labels in either dataset. We first learn a feature representation on the large dataset using deep convolutional embedded clustering (DCEC) [6], [9]. The motivation behind DCEC is that it simultaneously learns a feature representation and clusters the data on top of the feature representation. We use the DCEC cluster assignments as cluster labels on the annotated MOS dataset to perform multi-task learning without using the degradation information. The weights of the multi-task network are initialized with DCEC, similarly to the MTL approach above with the degradation classifier. We believe that using cluster labels might be beneficial for two reasons. First, the cluster assignments could represent concepts that are more complex than a degradation label. A speech signal is characterised by many factors which include rhythm, pitch, timbre and linguistic content. These factors are poorly represented by a degradation label. Instead, the cluster labels might represent a high-level semantic similarity between the data points and as a consequence classification of such cluster labels can be seen as a useful auxiliary task. Secondly, when transferring the weights from DCEC we might deteriorate the feature representation due to the optimization of the weights for the target task i.e., MOS prediction. Therefore, when doing multi-task learning with the output of DCEC we help the network to retain existing knowledge of the learnt representation from the large dataset.

SEMTL description: The proposed SEMTL-based model is represented in Fig. 2. In the first step, we use DCEC to cluster the large dataset. In the second step, we first assign clusters to the data in the small set by freezing the trained DCEC network. After annotating the dataset with the cluster assignments, we perform multi-task learning. The two tasks consist of cluster label classification and MOS prediction as follows:

$$\mathcal{L}_{tot} = \mathcal{L}_{ce}(Y^{(cl)}, \hat{Y}^{(cl)}) + \mathcal{L}_{mse}(S, \hat{S}) \tag{2}$$

where Y^{cl} and $\hat{Y^{cl}}$ represent respectively the cluster labels and their prediction, S is the MOS, $\mathcal{L}_{ce}$ is the cross-entropy and $\mathcal{L}_{mse}$ is the mean squared error. The multi-task network is initialized with the DCEC weights so that we use a learnt feature representation.

DCEC explained: DCEC consists of a convolutional autoencoder and a clustering layer that is attached to the embedded layer of the autoencoder. The embedded points z_i representing the input audio data are mapped by the clustering layer into a soft label using the Student's t-distribution:

$$q_{ij} = \frac{(1 + \|z_i - u_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j'}(1 + \|z_i - u_{j'}\|^2/\alpha)^{-\frac{\alpha+1}{2}}} \tag{3}$$

where q_{ij} is interpreted as the probability to assign the embedded point z_i to a cluster u_j. The parameter α is set to 1 in all the experiments. The number of clusters J is arbitrarily chosen and cluster centres are initialized by training K-means

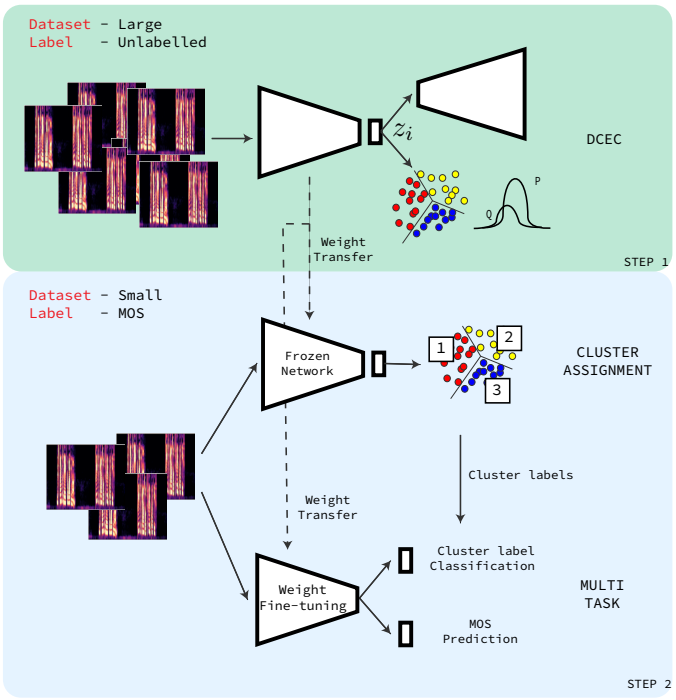


Fig. 2. SEMTL-based model. STEP 1 consists of performing DCEC on the large set. In STEP 2 we first extract the cluster assignments on the small dataset using DCEC, which has been previously trained in STEP 1. Then, the multi-task network is trained to classify the cluster assignments and to predict MOS simultaneously. The model in STEP 2 is initialized with the weights learned in STEP 1.

on the embedded features of the convolutional autoencoder. DCEC is based on minimizing two loss functions as follows:

$$\mathcal{L} = \mathcal{L}_r + \gamma \mathcal{L}_c \tag{4}$$

where $\mathcal{L}_r = |x - x'|_2^2$ is the cost function of the autoencoder with x representing the input data. The second term is a clustering loss defined as $\mathcal{L}_c = KL(P \parallel Q)$ which is the Kullback-Leibler divergence between the soft label assignments q_{ij} and an auxiliary target distribution p_{ij} defined as:

$$p_{ij} = \frac{q_{ij}^2/f_{ij}}{\sum_{j'}(q_{ij'}^2/f_{ij'})} \tag{5}$$

where $f_{ij} = \sum_i q_{ij}$. The auxiliary target distribution is based on giving importance to the most confident predictions and penalizing clusters with too many samples. As in [6] we fix γ to 0.1.

DCEC optimization: We optimize DCEC with convolutional autoencoders as done in [6]. We first train an autoencoder to initialize the DCEC network parameters and the cluster centres. Then we minimize the DCEC cost function which updates both cluster centres u_j and network parameters. All the learnable parameters can be updated with backpropagation as shown in [6], [9]. The auxiliary target distribution is updated every T iterations to avoid instability.

IV. Experiments

A. Datasets

TCD-VoIP [20] is used as the small MOS annotated dataset which consists of 384 recordings sampled at 48 kHz. This dataset size is typically considered as not sufficient for training an efficient deep learning model[1]. We take the audio stimuli with the following degradations: chop, clip, echo and background noise, collecting around 45 minutes of data. The dataset consists of degraded stimuli created with clean speech taken from the TSP speech database [25]. Speech sentences have a duration of ≈8 seconds and include 4 speakers (2 Male, 2 Female). TCD-VoIP includes several conditions for each degradation and there are 4 clips for each condition.

The large annotated dataset is built with the same degradations of TCD-VoIP using different speakers and sentences from the TSP database. TSP includes 24 adult speakers. We discarded the 4 speakers used in TCD-VoIP and we use the remaining 20 speakers. In this way, we make sure that no speaker or sentence dependent biases are transferred from the model trained on the large dataset to the one trained on TCD-VoIP. For each degradation, we include more conditions than the ones present in the TCD-VoIP to improve generalization in our model. We generate 3805 recordings which are almost 8.5h of audio. The dataset is divided into 761 stimuli per class. In total, we use 5 classes: CHOP, CLIP, ECHO, NOISE, REFERENCE. The reference speech is included as distinguishing degraded speech from clean speech could be useful for quality prediction.

B. Experiment set-up

Initial experiments showed that using 48 kHz sampling was not adding any benefit. Therefore, we downsample the data to 16 kHz to reduce the input dimension. We transform each raw audio waveform to log mel spectrograms using 64 mel bands and windows of 25 ms with 10 ms hop length. In both MTL and SEMTL we use 5 classes as we want to equally compare SEMTL with MTL. Therefore, we choose 5 clusters for DCEC and 5 classes for the auxiliary task in SEMTL (i.e., we classify 5 cluster labels in the auxiliary task). DCEC is trained in two steps. First, we train an autoencoder for 200 epochs. Secondly, DCEC with the clustering loss is trained until the number of cluster assignments between two consecutive auxiliary target distribution updates is lower than a threshold. We set the convergence threshold to 0.1% of the dataset size. We update the auxiliary target distribution every 70 batches. In [6] the target distribution is updated every 140 steps but our experiments showed instability. The supervised classifier is trained for 200 epochs as well. In all the experiments on the large dataset, we use a batch size of 64 and we update the weights using Adam optimizer with a learning rate of 0.001. The models on the TCD-VoIP dataset are trained using 10-fold cross-validation as the size is too small for splitting into training and test sets. In each fold, we train with $\frac{9}{10}$ of the data and test on the remaining $\frac{1}{10}$. We store the predicted quality score for each clip belonging to the test set in that fold. After training all the folds we have a predicted quality score for each clip in the TCD-VoIP dataset. The models trained on TCD-VoIP (i.e., the ones that predict quality scores) are optimized for 40 epochs in each fold, using Adam with learning rate 0.00001 and a batch size of 64.

[1]In the results section we show the performance of a naive baseline model trained on this dataset only.

TABLE I

Fully connected networks attached to the ConvNet based on the task type. When we classify degradations or cluster assignments we attach "Classification". When we predict MOS scores we attach the network "Regression". We use a shifted sigmoid σ as the final activation function to map the MOS range from 1 to 5. DCEC is composed by two fully connected layers, 10 neurons for the embedded layer and 5 neurons for the clustering layer. In each multi-task network, we attach both "Classification" and "Regression" to the same shared ConvNet.

Task	**Model**
DCEC	ConvNet $\rightarrow$ FC_{10} $\rightarrow$ FC_5
Classification	ConvNet $\rightarrow$ FC_{256} $\rightarrow$ ReLU $\rightarrow$ D(0.5) $\rightarrow$ FC_5 $\rightarrow$ Softmax
Regression	ConvNet $\rightarrow$ FC_{256} $\rightarrow$ ReLU $\rightarrow$ D(0.5) $\rightarrow$ FC_1 $\rightarrow$ $1+4\sigma$

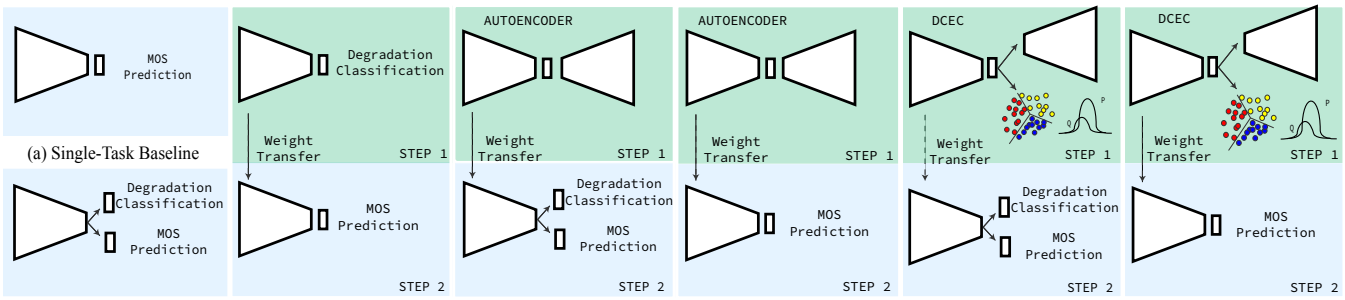
C. Architecture

We use the same convolutional architecture (ConvNet) in every model and we attach different fully connected layers depending on the task. Fixing the same ConvNet is required so that we can transfer the weights and we can fairly compare the different feature representations. The ConvNet consists of 4 layers $L^5_{32} \rightarrow L^5_{64} \rightarrow L^3_{128} \rightarrow L^3_{256} \rightarrow$ where L^k_m means a convolutional layer with m kernels and (k, k) kernel size. We used stride 2 and "same" padding in all the layers. This architecture represents the encoder in the autoencoder used in DCEC and the convolutional part in all of the other models. The structure of the decoder is the mirror of the encoder. In each layer, we use the ReLU activation function and batch normalization. For each task, we attach a fully connected network to the ConvNet as summarised in Table I.

D. Results

In our experiments, we want to compare MTL and SEMTL to each other. Also, we compare our proposed models with different combinations of feature representations with single-task or multi-task scenarios as shown in Fig. 3. We also explore the multi-task model using degradation classification as an auxiliary task after learning features with a convolutional autoencoder trained on the large dataset. The autoencoder is the same that we use to initialize DCEC.

We test the predicted scores against MOS using root-mean-square error (RMSE), Pearson correlation coefficient (PCC) and Spearman's rank-order correlation coefficient (SRCC) [26] as shown in Table II. The results are collected per condition. We predict quality scores for each clip and we compute the average of all the clips belonging to the same condition. Results show that every multi-task model combined with feature learning outperforms multi-task only



(a) Single-Task Baseline (b) Multi-Task Baseline (c) Degr. Single-Task (d) AE Multi-Task (e) AE Single-Task (f) DCEC Multi-Task (g) DCEC Single-Task

Fig. 3. Additional models that we used in the experiments. Step 1 is trained with the large dataset while step 2 with TCD-VoIP. In step 2 we do fine-tuning in each model. The two baselines (a) and (b) are end-to-end.

TABLE II
PERFORMANCE EVALUATION WITH ROOT MEAN SQUARE ERROR (RMSE), PEARSON CORRELATION COEFFICIENT (PCC), AND SPEARMAN'S RANK-ORDER CORRELATION COEFFICIENT (SRCC).

RMSE	**CHOP**	**CLIP**	**ECHO**	**NOISE**	**ALL**	**P.Sup23**
Single-Task Baseline	0.802	0.838	0.938	0.571	0.786	0.616
Multi-Task Baseline	0.778	0.754	0.926	0.533	0.753	/
Degr. Single-Task	**0.632**	0.599	0.766	0.391	0.605	0.571
AE Multi-Task	0.702	**0.489**	**0.570**	0.395	0.557	/
AE Single-Task	0.704	0.545	0.664	0.397	0.591	0.558
DCEC Multi-Task	0.676	**0.487**	**0.594**	**0.383**	**0.551**	/
DCEC Single-Task	0.688	0.524	0.665	0.405	0.584	**0.507**
MTL	0.684	0.561	0.628	0.358	0.569	/
SEMTL	0.678	0.491	0.595	**0.383**	**0.552**	**0.461**
P.563	**0.652**	0.627	0.975	0.945	0.827	1.119
SRMRnorm	0.936	1.023	1.023	0.982	0.986	0.696

PCC	**CHOP**	**CLIP**	**ECHO**	**NOISE**	**ALL**	**P.Sup23**
Single-Task Baseline	0.421	0.489	0.694	0.847	0.655	0.648
Multi-Task Baseline	0.487	0.686	0.780	0.878	0.712	/
Degr. Single-Task	**0.784**	0.894	0.809	0.924	0.812	0.813
AE Multi-Task	0.698	0.950	**0.907**	0.922	0.846	/
AE Single-Task	0.670	0.927	0.881	0.921	0.825	0.718
DCEC Multi-Task	0.701	**0.963**	0.894	**0.926**	**0.850**	/
DCEC Single-Task	0.673	0.943	0.876	0.917	0.831	**0.816**
MTL	0.720	0.911	**0.907**	**0.937**	0.837	/
SEMTL	0.703	**0.964**	0.894	**0.926**	**0.850**	**0.861**
P.563	**0.761**	0.887	0.712	0.356	0.637	0.664
SRMRnorm	0.482	0.694	0.640	0.634	0.576	0.616

SRCC	**CHOP**	**CLIP**	**ECHO**	**NOISE**	**ALL**	**P.Sup23**
Single-Task Baseline	0.440	0.437	0.596	0.827	0.606	0.680
Multi-Task Baseline	0.560	0.665	0.771	0.888	0.723	/
Degr. Single-Task	**0.825**	0.881	0.813	0.906	0.792	0.804
AE Multi-Task	0.696	**0.942**	0.855	0.911	**0.828**	/
AE Single-Task	0.633	0.920	**0.887**	**0.920**	0.811	0.723
DCEC Multi-Task	0.652	**0.942**	0.840	0.907	0.826	/
DCEC Single-Task	0.655	**0.956**	0.866	0.906	0.823	**0.811**
MTL	0.746	0.881	**0.906**	**0.912**	**0.827**	/
SEMTL	0.652	**0.942**	0.843	0.907	**0.827**	**0.853**
P.563	**0.786**	0.745	0.681	0.162	0.569	0.677
SRMRnorm	0.534	0.683	0.669	0.721	0.613	0.579

or pre-training only as well as the two baselines P.563 and SRMRnorm[2]. RMSE and PCC show that the multi-task models with features learnt by DCEC have the highest performance. SEMTL achieves the same results as DCEC Multi-Task with the advantage of not using the degradation labels, which is very encouraging. MTL is not able to outperform AE Multi-Task i.e., multi-task learning with feature learnt by an autoencoder.

The same procedure described in step 2 of each proposed model was repeated using the ITU-T P.Sup23 Experiment 1 database [27], which is a speech codec dataset including different languages (Japanese, French, and English). We test unseen degradations to evaluate whether the TCD-VoIP results are due to the usage of the same degradations between the larger dataset and the small one. Therefore, we only transferred the weights from the model trained on the same large dataset and we fine-tuned on P.Sup23 Exp1. Compared to the TCD-VoIP models, we only changed the learning rate to 0.0001. The results on the P.Sup23 (Table II) show that DCEC is robust on unseen degradations and unseen languages. Compared to TCD-VoIP, there is a larger gap between DCEC-based models (DCEC single-task and SEMTL-based model) and the autoencoder. This suggests that our proposed approach has better generalisation capacity than the autoencoder.

[2]P.563 and SRMRnorm are computed from https://github.com/qin/p.563 and https://github.com/MuSAELab/SRMRToolbox on data downsampled to 8 kHz.

E. Cluster analysis

DCEC clusters on TCD-VoIP are shown in Fig. 4 (a) against the degradation type and in Fig. 4 (b) against MOS. Both figures suggest that DCEC performs clustering according to a criterion that does not correspond to either degradations or MOS. Considering that DCEC learns a robust feature representation for quality prediction, it is plausible to assume that data points are grouped based on a high-level semantic similarity. Also, our results show that the simultaneous classification of the cluster assignments is beneficial for quality prediction which suggests that a meaningful grouping of the data points occurs. In Fig. 4 (b) we can see that subgroups based on MOS occur for NOISE only. This suggests that the different signal-to-noise ratio levels of the background noise are easily captured by DCEC. However, complex non-linear distortions such as CHOP, ECHO, and CLIP do not show this trend.

V. DISCUSSION

Multi-task learning combined with unsupervised feature representation learning shows promising results for non-intrusive speech quality prediction. RMSE and PCC (Table II) show that the unsupervised representation can replace a fully supervised degradation classifier.

The results of this paper suggest that going towards a better feature representation from unlabelled data is a promising approach and that might be taken into consideration as opposed to collecting large annotated MOS datasets which is not always affordable. We have also shown that our approach achieves very good generalization performance on unseen degradations

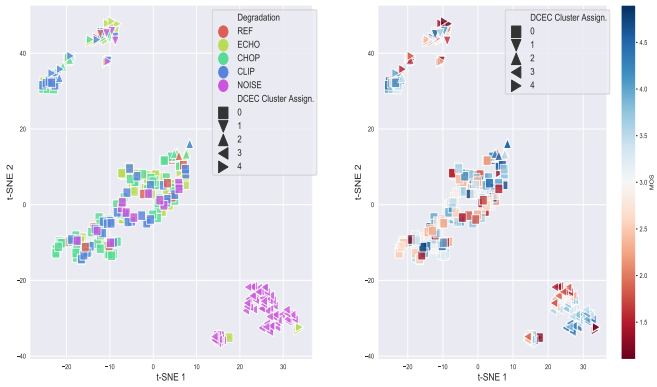


Fig. 4. Comparing DCEC cluster assignments with degradations (a) and MOS (b) using t-SNE. "REF" data points represent the first condition per each degradation which is clean speech as claimed by the TCD-VoIP authors [20].

and languages (P.Sup 23) compared to the autoencoder which is the only unsupervised feature learning approach proposed so far for speech quality assessment [18].

VI. Conclusions and Future Work

In this paper, we have proposed two multi-task learning approaches combined with unsupervised feature learning for non-intrusive speech quality assessment. In the MTL-based approach, we use degradation classification either as an auxiliary task and for learning the initial weights. The SEMTL-based approach consists of classifying cluster labels generated from DCEC, a deep clustering technique that learns features and clusters the data simultaneously. We have shown that multi-task learning combined with unsupervised feature learning shows promising performance for non-intrusive speech quality assessment using a very small MOS annotated dataset. In particular, SEMTL does not need any auxiliary labels and achieves the same performance as DCEC Multi-Task which uses degradation classification as an auxiliary task. Also, we have shown that DCEC learns good feature representations for speech quality prediction achieving higher RMSE and linear correlation than the autoencoder and the degradation classifier. Similar trends are shown on unseen degradations and languages.

In the future, we will evaluate this approach with a larger dataset (one order of magnitude bigger). We will design a transfer learning approach where we take a dataset with degradations from different applications (e.g., speech enhancement, audio codecs etc.). The experiments shown in this paper do not explore the number of clusters in DCEC. We believe that feature representation learned with DCEC might be sensitive to the number of clusters and that the optimal number of clusters could have not been found in this paper. Finally, we are aware that we have used a basic multi-task approach (e.g., we have not found optimal weight loss). We will explore different multi-task techniques with both degradation types and cluster assignments.

References

[1] ITU-T. P.800 Methods for subjective determination of transmission quality, 1996.

[2] Ludovic Malfait, Jens Berger, and Martin Kastner. P. 563—the ITU-T standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1924–1934, 2006.

[3] João F Santos, Mohammed Senoussaoui, and Tiago H Falk. An improved non-intrusive intelligibility metric for noisy and reverberant speech. In *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 55–59. IEEE, 2014.

[4] Babak Naderi, Rafael Zequeira Jiménez, Matthias Hirth, Sebastian Möller, Florian Metzger, and Tobias Hoßfeld. Towards speech quality assessment using a crowdsourcing approach: evaluation of standardized methods. *Quality and User Experience*, 6(1):1–21, 2020.

[5] Tobias Hoßfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia. Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *IEEE Transactions on Multimedia*, 16(2):541–558, 2013.

[6] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. Deep clustering with convolutional autoencoders. In *International conference on neural information processing*, pages 373–382. Springer, 2017.

[7] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines. Development of a speech quality database under uncontrolled conditions. In *Proc. Interspeech 2020, pp. 4616–4620*, 2020.

[8] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines. Adapting the quality of experience framework for audio archive evaluation. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.

[9] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016.

[10] Anderson R Avila, Hannes Gamper, Chandan Reddy, Ross Cutler, Ivan Tashev, and Johannes Gehrke. Non-intrusive speech quality assessment using neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 631–635. IEEE, 2019.

[11] Benjamin Cauchi, Kai Siedenburg, Joao F Santos, Tiago H Falk, Simon Doclo, and Stefan Goetze. Non-intrusive speech quality prediction using modulation energies and LSTM-Network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(7):1151–1163, 2019.

[12] Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang. Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM. In *Proc. Interspeech 2018, pp. 1873-1877*, 2018.

[13] Xuan Dong and Donald S Williamson. A classification-aided framework for non-intrusive speech quality assessment. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 100–104. IEEE, 2019.

[14] Andrew A Catellier and Stephen D Voran. Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 331–335. IEEE, 2020.

[15] Gabriel Mittag and Sebastian Möller. Non-intrusive speech quality assessment for super-wideband speech communication networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7125–7129. IEEE, 2019.

[16] Jasper Ooster, Rainer Huber, and Bernd T. Meyer. Prediction of perceived speech quality using deep machine listening. In *Proc. Interspeech 2018*, pages 976–980, 2018.

[17] Joan Serrà, Jordi Pons, and Santiago Pascual. SESQA: semi-supervised learning for speech quality assessment. *arXiv preprint arXiv:2010.00368*, 2020.

[18] Meet H Soni and Hemant A Patil. Novel deep autoencoder features for non-intrusive speech quality assessment. In *European Signal Processing Conference (EUSIPCO)*, pages 2315–2319. IEEE, 2016.

[19] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

[20] Naomi Harte, Eoin Gillen, and Andrew Hines. TCD-VoIP, a research database of degraded speech for assessing quality in VoIP applications. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2015.

[21] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[22] Le Kang, Peng Ye, Yi Li, and David Doermann. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *2015 IEEE international conference on image processing (ICIP)*, pages 2791–2795. IEEE, 2015.

[23] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[24] Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*, pages 561–580. Springer, 2012.

[25] Peter Kabal. TSP speech database. *McGill University, Database Version*, 1(0):09–02, 2002.

[26] ITU-T. P.1401 Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models., 2020.

[27] ITU-T. P. Supplement 23 ITU-T coded-speech database, 1998.