INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Probing Speech Quality Information in ASR Systems

Bao Thang Ta, Minh Tu Le, Nhat Minh Le, Van Hai Do

Viettel Cyberspace Center, Viettel Group, Hanoi, Vietnam
{thangtb3, tulm7, minhnl2, haidv21}@viettel.com.vn

Abstract

This paper investigates how intermediate speech representations in a state-of-the-art automatic speech recognition (ASR) system encode multi-dimensional speech quality, including MOS, Noisiness, Coloration, Discontinuity, and Loudness. We found that speech quality information is encoded in the ASR encoder layers at various levels but is still much richer than the Mel-spectrogram, an input widely used in previous works. This discovery inspires us to develop the Attentive Conformer with ASR pretraining, a novel deep learning model that enables the utilization of rich information from pretrained ASR models and the ability to focus on specific layers. Experiments on the NISQA speech quality assessment dataset demonstrate that the proposed model achieves state-of-the-art performance with significantly less training data.
Index Terms: Speech Quality Assessment, MOS, Speech Recognition, Conformer, Transfer Learning

1. Introduction

Speech quality assessment is a crucial task in evaluating the performance of communication systems. It is imperative to ensure that speech transmitted over networks meets the required quality standards. While the subjective method is the most straightforward way to assess speech quality, it is expensive and time-consuming, making it impractical for real-time and large-scale systems. Objective methods have been developed to overcome these limitations [1]. Among objective methods, single-ended methods have gained significant attention because they do not require a clean or reference speech signal, unlike double-ended methods such as PESQ [2] and POLQA [3]. The use of the double-ended method is limited because reference speech signals are usually unavailable in most realistic scenarios. Therefore, the development of single-ended methods has become an active research area in speech quality assessment.

In recent years, deep learning-based methods have shown remarkable performance in single-ended speech quality assessment. Various neural network architectures have been explored with different features extracted from speech signals. For instance, an autoencoder was used to extract spectrogram information in [4], while CNN was implemented with waveform input [5] or Q spectrum [6]. Furthermore, a modulation energy-based LSTM network was proposed in [7], while a CNN and LSTM model combination was presented in [8]. The use of self-attention with CNN was proposed in [9]. These studies have shown promising results in predicting speech quality and provide valuable insights for developing effective models for speech quality assessment.

However, the limited availability of datasets for speech quality assessment is a significant challenge that hinders the development of effective models in this field [9]. The process of collecting subjective ratings is time-consuming, expensive, and requires controlled experimental conditions, which makes it difficult to gather a significant number of reliable and diverse ratings. Recently, some studies proposed using self-supervised models such as HuBERT and wav2vec2 [10, 11, 12], which were trained on large amounts of unlabeled audio data to learn general-purpose feature representations of speech. These models have been used as feature extractors or finetuned for speech quality assessment, allowing for the creation of models with high performance despite limited datasets. Despite these benefits, it is noted that these models may not be optimized for specific tasks such as speech quality assessment, and they often require significant computational resources, which can be challenging to obtain for researchers and organizations with limited resources. Additionally, deploying these models in resource-constrained environments can be difficult due to their large size.

An alternative approach is to utilize pre-trained Automatic Speech Recognition (ASR) models, which have made many advances in recent years [13]. These models can capture speech's complex linguistic and acoustic characteristics while having a much smaller size than wav2vec2 and HuBERT, making them a viable option for speech quality assessment. Furthermore, ASR systems are trained to recognize spoken words and transcribe them into text, requiring them to learn how to process various aspects of speech, such as pitch, timing, and spectral features, which are also relevant to perceived speech quality.

However, no previous research has explored the potential of pre-trained ASR models for speech quality assessment, despite their demonstrated effectiveness in several speech-related tasks, such as accent recognition [14, 15, 16], emotion classification [17, 18, 19], speaker verification [20], and keyword spotting [21]. This study addresses this gap by investigating the potential of pre-trained ASR models for speech quality assessment tasks. By leveraging the knowledge acquired through ASR tasks, we believe that speech quality assessment models can achieve high performance even with smaller datasets, making it a cost-effective and efficient approach to developing effective models for speech quality assessment.

To obtain a state-of-the-art ASR model for analysis, we use Conformer [13] as the core of our ASR encoder. Through probing experiments, we found that the information embedded by different ASR encoder layers can significantly improve prediction results compared to directly using Mel-spectrogram as input, which has been widely used in previous works [7, 8, 22, 9]. Besides, we design a novel attention mechanism to combine information from various ASR encoder layers into a high-performance multitask speech quality assessment model, which predicts not only MOS but also other subjective dimensions: Noisiness (NOI), Coloration (COL), Discontinuity (DIS), and

10.21437/Interspeech.2023-2507

Loudness (LOUD). The comparisons with recent state-of-the-art models NISQA [9], wav2vec2 [10] and several modern deep learning models point out the significant improvement of our proposal while using a smaller amount of training data.

This paper is structured as follows. Section 2 explains the proposed speech quality assessment model and how the speech quality information is probed from each ASR encoder layer. The experimental results confirming the proposed approach's effectiveness are presented in Section 3. Finally, in Section 4, we summarize our research findings and outline the future directions for this work.

2. Proposed Method

2.1. Probing Speech Quality Information in ASR layers

Figure 1 demonstrates how we investigate MOS information in ASR encoder layers. To explore the speech quality information stored in ASR encoder layers, we employed an end-to-end 15-layer ASR model based on the advanced Conformer architecture [13]. The model was trained on a substantial amount of labeled data from the Librispeech datasets [23], ensuring its ability to gather extensive and diverse information from the available speech samples. We then froze all its encoder layers and utilized them as a feature extractor for speech quality assessment tasks. For each layer, we pass its embedding

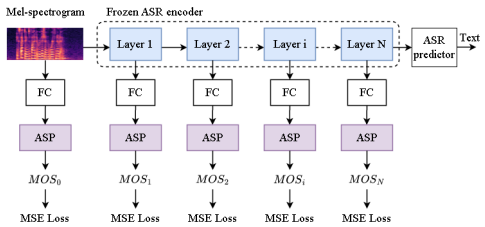

Figure 1: *Probing speech quality information in ASR layers.*

through a Fully Connected (FC) layer, an Attentive Statistical Pooling (ASP) layer, and a sigmoid function. By using the ASP method, we obtain global information about the entire utterance while paying attention to unique frames. The mathematical formula for the proposed probing method is as follows: Given an input embedding at the i^{th} layer denoted $H = [h_1, h_2, \ldots, h_T] \in \mathbb{R}^{T \times D}$ with $h_t = [h_{t1}, h_{t2}, \ldots, h_{tD}]$ where T is the number of frames, and D is the embedding size, the speech quality information is evaluated as:

$$\begin{aligned} h'_t &= FC(h_t); e_t = v^T f(W h'_t + b) \\ \alpha_t &= \frac{e_t}{\sum_{k=1}^{T} e_k}; \mu = \sum_{t=1}^{T} \alpha_t h'_t \\ \sigma &= \sqrt{\sum_{t=1}^{T} \alpha_t h'_t \odot h'_t - \mu \odot \mu} \\ x &= [\mu, \sigma] \\ MOS_i &= 5 \times f_1(W_1 x + b_1) \end{aligned} \tag{1}$$

where $v \in \mathbb{R}^D$ is the weight vector, $W \in \mathbb{R}^{D \times D}$ and $W_1 \in \mathbb{R}^{2D \times 1}$ are the weight matrices, $b \in \mathbb{R}^D$ and $b_1 \in \mathbb{R}^1$ are bias items, and $f(.)$ is a non-linear activation function, such as RELU. $\mu \in \mathbb{R}^D$ and $\sigma \in \mathbb{R}^D$ are the weighted mean vector and weighted standard deviation over all frames, respectively. f_1 is a sigmoid function. In a similar manner, we perform probing models for other speech-quality indicators: NOI, COL, DIS, and LOUD. All probing models are trained with the same learning rate of 0.001, batch size of 16, and 200 epochs on the same computer.

Probing experiments are conducted on four datasets, two for training (NISQA_TRAIN_SIM and NISQA_TRAIN_LIVE) and two for validation (NISQA_VAL_SIM and NISQA_VAL_LIVE), sourced from the NISQA corpus [9]. The datasets NISQA_TRAIN_SIM and NISQA_VAL_SIM were created by simulating various speech distortions, such as packet loss and clipping, while NISQA_TRAIN_LIVE and NISQA_VAL_LIVE consisted of live recordings with genuine distortions like typing on a keyboard and street noise. All datasets were annotated using ITU-T P.808 [24] with five ratings per file. Before training and validation, the samples were down-sampled to 16 kHz to be compatible with the pretrained ASR model. Pearson's correlation coefficient (PCC) was used as the evaluation metric to measure the correlation between predicted and subjective speech quality values.

The results, depicted in Figure 2, show the PCC between predicted and subjective speech quality values at different encoder layers of the pretrained ASR model. The findings indicate that different layers store varying amounts of information about speech quality, with the last layers fading the speech quality information more than the earlier layers. However, using embeddings from ASR encoder layers (e.g., layers 1 to 15) still helps the model predict much better quality information than Mel-spectrogram (layer 0), especially for the discontinuity indicator. This finding indicates that speech-quality information is well encoded in ASR encoder layers and exploring this information could be beneficial for speech quality assessment tasks.

2.2. Proposed Speech Quality Assessment Model

The above analysis highlights the benefits of using information stored in trained ASR encoder layers for speech quality assessment tasks. However, the amount of information varies across layers and tends to decrease in the last layers due to their dependence on the ASR task. For example, some information relevant to speech quality, such as noise, is suppressed in the last layers. To reduce this dependence, we propose fine-tuning the pretrained models rather than using them solely as feature extractors. This allows for updates to the weights in the pretrained ASR encoder along with all the other weights in the proposed model. Secondly, to account for the varying amount of information stored across layers, instead of only using information at the last layer, we recommend using the attention mechanism on top of the ASR layers to generate a global quality score for each speech quality indicator, as illustrated in Figure 3. Furthermore, to enhance computational efficiency, we develop a multitask model that predicts all indicators (MOS, NOI, COL, DIS, and LOUD) simultaneously, sharing an encoder across tasks.

Our multitask model is trained using five constituent losses: MOS, NOI, COL, DIS, and LOUD loss. In contrast to prior approaches that rely on manually tuning or using equal loss weights, we employ an uncertainty loss that uses the homoscedastic uncertainty of each task to weigh multiple loss functions, as proposed in [25]:

$$\begin{aligned} \mathcal{L} = &\frac{\mathcal{L}_{MOS}}{\sigma^2_{MOS}} + \frac{\mathcal{L}_{NOI}}{\sigma^2_{NOI}} + \frac{\mathcal{L}_{COL}}{\sigma^2_{COL}} + \frac{\mathcal{L}_{DIS}}{\sigma^2_{DIS}} + \frac{\mathcal{L}_{LOUD}}{\sigma^2_{LOUD}} \\ &+ \log(\sigma^2_{MOS}\sigma^2_{NOI}\sigma^2_{COL}\sigma^2_{DIS}\sigma^2_{LOUD}) \end{aligned} \tag{2}$$

where σ_{MOS}, σ_{NOI}, σ_{COL}, σ_{DIS}, and σ_{LOUD} are learnable

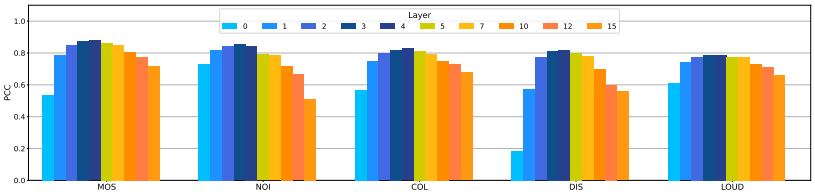

Figure 2: *Averaged PCC between predicted and subjective speech quality values in ASR encoder layers on two validation datasets (NISQA_VAL_SIM and NISQA_VAL_LIVE). For layer 0, the probing model directly uses the Mel-spectrogram as input.*

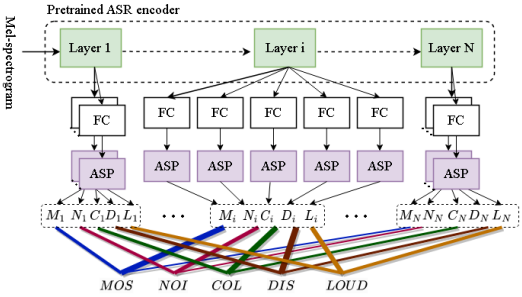

Figure 3: *Proposed Speech Quality Assessment Model.*

parameters. $\mathcal{L}_A$ represents the Mean Squared Error (MSE) between predicted and subjective quality values of the speech quality indicator A, where A can take values of MOS, NOI, COL, DIS, or LOUD.

3. Experiments

3.1. Setup

Our quality assessment model has 15 Conformer layers with an embedding size of 320, 4 attention heads, a convolution kernel size of 31, and other setting consistent with the medium Conformer [13]. It has about 39.8 million parameters. The input features are 80-dimensional Mel-spectrograms with a window of 25 ms and a frame shift of 10 ms, normalized using cepstral mean subtraction [26]. The model is trained for 300 epochs with a learning rate of 0.001 and a batch size of 16.

3.2. Datasets

We adopt only two datasets which are NISQA_TRAIN_SIM and NISQA_TRAIN_LIVE, for the training process. The validation and testing are conducted on five datasets: NISQA_VAL_LIVE, NISQA_VAL_SIM, NISQA_TEST_FOR, NISQA_TEST_P501, and NISQA_TEST_LIVETALK. All these datasets are obtained from the NISQA corpus [9]. The details of these datasets are shown in Table 1.

Table 1: *The description of used datasets*

Datasets	Source	#Files	Hours
NISQA_TRAIN_SIM	AusTalk [27], TSP [28]	10,000	24.7
NISQA_VAL_SIM	DNS Challenge [29], UK-Ireland [30]	2,500	6.0
NISQA_TRAIN_LIVE	Live phone and Skype	1,020	2.6
NISQA_VAL_LIVE		200	0.5
NISQA_TEST_FOR	Forensic speech dataset	240	0.6
NISQA_TEST_LIVETALK	Real phone and VoIP calls	232	0.6
NISQA_TEST_P501	ITU-T Rec. P.501	240	0.5

3.3. Scenario

We assess the proposed model with several baselines:

- Several modern models, including ResNet34 [31] (24M parameters), ECAPA-TDNN [26] (42M parameters), and Conformer [13] (39.8M parameters), are used as baselines. ECAPA-TDNN was configured with a parameter approximation to our proposed model, while Resnet34 was used with its default configuration. The residual block channels in ResNet34 are set as {64, 128, 256, 512}, while the SE-Res2Blocks channels in ECAPA-TDNN are {2048, 2048, 2048}. The Conformer model uses the same configuration as the proposed model. The models are trained from scratch without pretraining.
- Secondly, we consider two versions of a recent state-of-the-art speech quality assessment model, NISQA [9], named NISQA2 and NISQA59. These two versions share the same architecture but were trained with different amounts of training data. NISQA2 was retrained based on the official code using only two training datasets: NISQA_TRAIN_LIVE and NISQA_TRAIN_SIM, which contained a total of 11,020 speech files. Meanwhile, NISQA59 is the pretrained model provided by [9], which used 72,903 speech files from two of our training datasets and an additional 57 private datasets.
- Finally, we compare using pretrained ASR models to self-supervised models such as wav2vec2 (95M parameters) proposed by Cooper et al [10]. We fine-tune wav2vec2 only on two training datasets (NISQA_TRAIN_LIVE and NISQA_TRAIN_SIM) using the same loss function as our proposed model to ensure a fair comparison.

We utilize two criteria, Pearson's Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE), to evaluate the performance of the models.

3.4. Results and Discussion

The results of different models on each speech quality criterion for the NISQA_VAL_SIM and NISQA_VAL_LIVE datasets are presented in Tables 2 and 3. Additionally, Table 4 shows the average PCC and RMSE of models across all speech quality criteria for each dataset. Based on the findings in Tables 2 and 3, the Conformer model trained from scratch (row 6) outperforms the NISQA2 model (row 3) in terms of RMSE values on most speech quality criteria, except NOI. However, it performs worse in PCC values. When compared to contemporary models such as ECAPA-TDNN and ResNet34, the trained-from-scratch Conformer performs better in both PCC and RMSE values.

The use of pretrained ASR models leads to higher correlation and less error between predicted and subjective speech quality values, as shown in the last two rows of Tables 2, 3, and 4. The Conformer model using the pretrained ASR model (row 7) outperforms the trained-from-scratch Conformer model

Table 2: *PCC and RMSE of compared models on the NISQA_VAL_SIM dataset. Abbreviation:* ***Attn*** *- the proposed attention to combine information from all pretrained layers. NISQA59 is trained on 59 datasets, while other methods are trained only on two datasets.*

	MOS		NOI		COL		DIS		LOUD	
Model	**PCC ↑**	**RMSE ↓**	**PCC ↑**	**RMSE ↓**	**PCC ↑**	**RMSE ↓**	**PCC ↑**	**RMSE ↓**	**PCC ↑**	**RMSE ↓**
Resnet34	0.802	0.669	0.838	0.519	0.748	0.628	0.661	0.763	0.719	0.580
ECAPA-TDNN	0.796	0.689	0.849	0.509	0.740	0.638	0.643	0.791	0.735	0.573
NISQA2	0.896	0.532	0.871	0.486	0.835	0.581	0.823	0.619	**0.807**	0.533
wav2vec2	0.906	0.561	0.874	0.497	0.835	0.544	0.827	0.657	0.804	**0.507**
NISQA59	0.897	0.523	0.862	0.501	0.809	0.572	0.823	0.606	0.797	0.523
Conformer (from scratch)	0.880	0.528	0.860	0.501	0.816	0.544	0.797	0.604	0.767	0.533
Conformer (pretrained)	0.922	**0.440**	0.867	0.478	0.851	**0.519**	0.867	0.503	0.794	0.518
Conformer (pretrained) + Attn	**0.927**	0.458	**0.878**	**0.477**	**0.860**	0.523	**0.881**	**0.496**	0.799	0.527

Table 3: *PCC and RMSE of compared models on the NISQA_VAL_LIVE dataset*

	MOS		NOI		COL		DIS		LOUD	
Model	**PCC ↑**	**RMSE ↓**	**PCC ↑**	**RMSE ↓**	**PCC ↑**	**RMSE ↓**	**PCC ↑**	**RMSE ↓**	**PCC ↑**	**RMSE ↓**
Resnet34	0.720	0.493	0.601	0.588	0.444	0.519	0.503	0.607	0.640	0.537
ECAPA-TDNN	0.717	0.531	0.639	0.562	0.426	0.541	0.417	0.681	0.669	0.535
NISQA2	0.804	0.456	0.706	0.524	0.532	0.511	0.562	0.607	0.707	0.528
wav2vec2	**0.870**	0.441	0.724	0.506	**0.577**	**0.455**	0.547	0.660	**0.731**	**0.473**
NISQA59	0.822	0.401	0.723	0.550	0.566	**0.454**	0.542	0.604	0.728	0.492
Conformer (from scratch)	0.805	0.419	0.681	0.531	0.505	0.458	0.523	0.581	0.705	0.489
Conformer (pretrained)	0.859	**0.378**	0.749	0.489	0.533	0.500	0.598	0.556	0.728	0.478
Conformer (pretrained) + Attn	0.861	**0.378**	**0.761**	**0.477**	0.519	0.484	**0.618**	**0.555**	0.718	0.496

Table 4: *Averaged PCC and RMSE over all speech quality indicators of compared models*

Model	**NISQA_VAL_LIVE**		**NISQA_VAL_SIM**		**NISQA_TEST_FOR**		**NISQA_TEST_LIVETALK**		**NISQA_TEST_P501**	
	$\overline{\text{PCC}}$ ↑	**$\overline{\text{RMSE}}$ ↓**	**$\overline{\text{PCC}}$ ↑**	**$\overline{\text{RMSE}}$ ↓**	**$\overline{\text{PCC}}$ ↑**	**$\overline{\text{RMSE}}$ ↓**	**$\overline{\text{PCC}}$ ↑**	**$\overline{\text{RMSE}}$ ↓**	**$\overline{\text{PCC}}$ ↑**	**$\overline{\text{RMSE}}$ ↓**
Resnet34	0.582	0.549	0.754	0.632	0.650	0.639	0.664	0.692	0.731	0.679
ECAPA-TDNN	0.574	0.570	0.753	0.640	0.705	0.602	0.654	0.671	0.742	0.723
NISQA2	0.662	0.525	0.846	0.550	0.865	0.451	0.684	0.829	0.881	0.542
wav2vec2	0.690	0.507	0.849	0.553	0.863	0.447	0.778	0.650	0.834	0.666
NISQA59	0.676	0.500	0.838	0.545	0.866	0.528	0.702	0.725	0.877	**0.409**
Conformer (from scratch)	0.644	0.496	0.824	0.542	0.809	0.517	0.693	0.733	0.823	0.627
Conformer (pretrained)	0.693	0.480	0.860	**0.492**	**0.889**	0.386	0.769	0.641	0.874	0.525
Conformer (pretrained) + Attn	**0.695**	**0.478**	**0.869**	0.496	0.888	**0.385**	**0.795**	**0.609**	**0.884**	0.543

(row 6) in all 5 datasets, with an average relative improvement in PCC ranging from 4.2% to 9.9% and reduced RMSE by 3.1% to 16.3%.

Furthermore, the pretrained ASR-based Conformer model (row 7), which used only 2 datasets, also outperforms the NISQA59 model (row 5) trained on 59 datasets, on 4 out of 5 datasets (except NISQA_TEST_P501), as shown in Table 4. This result demonstrates the effectiveness of leveraging pretrained ASR models to achieve state-of-the-art performance with less training data.[1]

Additionally, as shown in Table 4, our approach (row 8), which uses the proposed attention method to combine information from all pretrained ASR layers, as described in Figure 3, exhibits superior PCC values compared to the model that solely utilizes information at the last layer (row 7) on 4 out of 5 datasets, and comparable performance on the remaining dataset. This serves as evidence of the efficacy of our proposed attention method.

Besides, Table 4 also indicates that compared to finetuning wav2vec2 (row 4), our approach (row 8) gives better PCC values on 5 out of 5 datasets. This indicates the competitiveness of using pretrained ASR models versus pretrained self-supervised models in speech quality assessment tasks.

Finally, to compare the inference speed of the models, we calculate their inverse of real-time factors (RTFX) on the NISQA_VAL_LIVE dataset, as shown in Table 5. The Conformer model has the highest speed with an RTFX of 89.6, while ResNet34 was the slowest with an RTFX of 31.9. The ECAPA-TDNN and NISQA models have medium speeds, with RTFXs of 66.3 and 49.9, respectively. The wav2vec2-based model only has an inference speed better than ResNet34. Interestingly, the Conformer model has a large number of parameters but still performs the fastest. In contrast, the ResNet34 model has fewer parameters, yet it is the slowest among the models. The reason may be that residual connections in its architecture add the input of a layer to the output of a later layer, increasing the depth of the network. Incorporating the proposed attention method slightly reduces the speed of our proposed model, but it remains much faster than NISQA, ECAPA-TDNN, wav2vec2, and ResNet34.

Table 5: *Inference speed of compared models on the same computer using a single CPU core (Intel Xeon Gold 5218R)*

Model	**RTFX ↑**
ResNet34	31.9
ECAPA-TDNN	66.3
NISQA	49.9
wav2vec2	43.4
Conformer (from scratch)	**89.6**
Conformer (pretrained)	**89.6**
Conformer (pretrained) + Attn	70.7

4. Conclusions

In this paper, we present a novel approach that leverages pretrained ASR models to predict multiple quality criteria in speech quality assessment tasks. Our approach outperforms several state-of-the-art approaches on various speech quality criteria, despite using less training data. Additionally, we show that the ASR encoder layers contain valuable speech-quality information, opening new opportunities for further development in the field of speech quality assessment.

[1] Since these 57 additional datasets are not public, we do not report the results of other models when using the same amount of data as NISQA59.

5. References

[1] P. C. Loizou, "Speech quality assessment," *Multimedia analysis, processing and communications*, pp. 623–654, 2011.

[2] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[3] ITU-T Recommendation P.863, "Perceptual objective listening quality assessment," 2011.

[4] M. H. Soni and H. A. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 2315–2319.

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

[6] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 631–635.

[7] B. Cauchi, K. Siedenburg, J. F. Santos, T. H. Falk, S. Doclo, and S. Goetze, "Non-intrusive speech quality prediction using modulation energies and LSTM-network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1151–1163, 2019.

[8] G. Mittag and S. Möller, "Quality degradation diagnosis for voice networks-estimating the perceived noisiness, coloration, and discontinuity of transmitted speech." in *INTERSPEECH*, 2019, pp. 3426–3430.

[9] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *INTERSPEECH*, pp. 2127–2131, 2021.

[10] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8442–8446.

[11] W.-C. Tseng, C. yu Huang, W.-T. Kao, Y. Y. Lin, and H. yi Lee, "Utilizing Self-Supervised Representations for MOS Prediction," in *Proc. Interspeech 2021*, 2021, pp. 2781–2785.

[12] H. Becerra, A. Ragano, and A. Hines, "Exploring the influence of fine-tuning data on wav2vec 2.0 model for blind speech quality prediction," in *Proc. Interspeech 2022*, 2022, pp. 4088–4092.

[13] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[14] B. T. Ta, X. V. Dang, Q. T. Duong, N. M. Le *et al.*, "Improving vietnamese accent recognition using asr transfer learning," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2022, pp. 1–6.

[15] A. Prasad and P. Jyothi, "How accents confound: Probing for accent information in end-to-end speech recognition systems," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3739–3753.

[16] Z. Zhang, Y. Wang, and J. Yang, "Accent recognition with hybrid phonetic features," *Sensors*, vol. 21, no. 18, p. 6258, 2021.

[17] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer Learning for Improving Speech Emotion Classification Accuracy," in *Proc. Interspeech 2018*, 2018, pp. 257–261.

[18] N. Tits, K. El Haddad, and T. Dutoit, "ASR-based features for emotion recognition: A transfer learning approach," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 48–52.

[19] B. T. Ta, T. L. Nguyen, D. S. Dang, N. M. Le *et al.*, "Improving speech emotion recognition via fine-tuning asr with speaker information," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1–6.

[20] S. Chen, Y. Wu, C. Wang, S. Liu, Z. Chen, P. Wang, G. Liu, J. Li, J. Wu, X. Yu, and F. Wei, "Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition?" in *Proc. Interspeech 2022*, 2022, pp. 3699–3703.

[21] D. Seo, H.-S. Oh, and Y. Jung, "Wav2KWS: Transfer learning from speech representations for keyword spotting," *IEEE Access*, vol. 9, pp. 80 682–80 691, 2021.

[22] X. Shu, Y. Chen, C. Shang, Y. Zhao, C. Zhao, Y. Zhu, C. Huang, and Y. Wang, "Non-intrusive speech quality assessment with a multi-task learning based subband adaptive attention temporal convolutional neural network," *Proc. Interspeech 2022*, pp. 3298–3302, 2022.

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

[24] ITU-T Recommendation P.808, "Subjective evaluation of speech quality with a crowdsourcing approach," 2018.

[25] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[26] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[27] D. Burnham, D. Estival, S. Fazio, J. Viethen, F. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner, Y. Kinoshita, R. Göcke, J. Arciuli, M. Onslow, T. Lewis, A. Butcher, and J. Hajek, "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable black box," in *Proc. Interspeech 2011*, 2011, pp. 841–844.

[28] P. Kabal, "Tsp speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[29] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech 2020*, 2020, pp. 2492–2496.

[30] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source multi-speaker corpora of the english accents in the british isles," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6532–6541.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.