ADTMOS – Synthesized Speech Quality Assessment Based on Audio Distortion Tokens

Qiao Liang, Ying Shen, Tiantian Chen, Lin Zhang, *Senior Member, IEEE*, Shengjie Zhao, *Senior Member, IEEE*

***Abstract*—In the fields of voice conversion (VC) and text-to-speech (TTS), recent years have witnessed a growing interest in developing synthesized speech quality assessment (SQA) systems. For such systems, it is essential to reliably and accurately evaluate the quality of synthesized speech produced by VC and TTS systems, as this remains a crucial issue requiring further exploration. Among various evaluation standards of speech quality, the mean opinion score (MOS) is the most commonly used SQA metric. The rapid advancement of deep learning (DL) techniques has propelled the emergence of DL-based MOS-based SQA algorithms. Unfortunately, none of these methods incorporate listeners' perceptions of audio distortions, which are considered one of the key factors affecting listeners' MOS ratings. To fill such a research gap to some extent, we propose a novel speech quality assessment framework, namely ADTMOS (Audio Distortion Token-Guided Deep MOS Predictor). ADTMOS consists of three parts: a public encoding layer which encodes the audio embeddings, an audio distortion token extractor which extracts ADT scores related to the subjective perceptions of audio distortions, and a frame-wise MOS score generator which is responsible for computing frame-level MOS scores. Experimental results demonstrate that compared to the LDNet baseline, ADTMOS achieves a 0.83% improvement on the VCC2018-CSMSC dataset and a 4.58% increase on the BVCC dataset in the system-level Spearman's rank correlation coefficient (SRCC). Furthermore, two innovative data augmentation techniques have been developed for the SQA task, aiming to mitigate the challenges of data scarcity and uneven sample distribution commonly encountered in SQA datasets. The source code is available at https://github.com/redifinition/ADTMOS.**

***Index Terms*—Speech quality assessment, MOS prediction, deep neural networks, speech enhancement, no-reference, audio distortions**

I. Introduction

With [1] the rapid development of text-to-speech (TTS) and voice conversion (VC) systems, speech quality assessment (SQA), which aims to reliably and accurately evaluate the quality of synthesized speech produced by those systems, has become increasingly important. SQA can be performed manually by human listeners (subjective assessment) or automatically by computational approaches (objective assessment). In subjective assessment, human listeners are recruited to evaluate the qualities of the synthesized speech. Then, the quality scores are given by listeners following specific evaluation standards, such as mean opinion score (MOS) [2], comparison mean opinion score (CMOS) [3], degradation mean opinion score (DMOS) [4], diagnostic acceptability measure (DAM) [2]. Among these, MOS is the most widely used SQA evaluation standard. Although it has limitations when used as an SQA metric, e.g., it is subjective and relies heavily on the listener's subjective feelings [5], and it can not be directly compared from different periods [6], it is still an important metric which reveals the speech quality from perspectives of clarity and naturalness.

The biggest problem with subjective assessment of MOS is that it is labor-intensive and time-consuming, and becomes impractical when a large number of speech samples need to be assessed. A typical solution is to randomly select a subset of samples, but this approach may yield unreliable results. As a result, automatic SQA, which uses computational methods to quickly and accurately predict MOS values of audios, has gained significant attention. Most automatic SQA models predict MOS scores that closely align with human subjective ratings, providing an efficient alternative to manual evaluation.

Automatic SQA approaches, also called objective assessment, can be classified into two categories, i.e., full-reference methods and no-reference methods [7]. Full-reference methods [8]–[13] assess the quality of synthesized speech based on provided reference samples. They can only work under the supervision of clear reference speech with the same content, duration, sample rate, number of channels, etc., which severely restricts their applications in practice. By contrast, no-reference methods evaluate the quality of synthesized speech without any reference speech, and consequently, they can be applied in broader scenarios.

No-reference methods can be further classified into two categories: rule-based and deep learning (DL)-based methods. The former ones adopt different techniques, such as vocal-tract modeling techniques [14], [15], temporal envelope representation technique [16], modulation spectral representation [17], and Gaussian Mixture Model (GMM) [18], to construct rules for SQA. However, these methods rely on manually designed features or prior knowledge, limiting them to evaluating specific distortion types, such as frequency and phase distortion. To address these limitations, deep learning (DL) techniques have been introduced in the past decade to improve assessment capabilities.

DL-based SQA models utilize audio features [19]–[23] or cross-domain features, including listener-dependent (LD) features [24], [25] and audio metadata [26], etc., to extract speech quality-related information. These features are fed into deep neural networks (DNNs) to produce speech-quality-related

Q. Liang, Y. Shen, T. Chen, L. Zhang and S. Zhao are with the School of Software Engineering, Tongji University, Shanghai 200082, China. E-mail: 2333091@tongji.edu.cn, yingshen@tongji.edu.cn, 2111287@tongji.edu.cn, cslinzhang@tongji.edu.cn, shengjiezhao@tongji.edu.cn. *(Corresponding author: Ying Shen.)*

embeddings. Several studies have also utilized fine-tuned self-supervised learning (SSL) models [27]–[31] to better leverage different audio features and achieve satisfied performances.

Previous DL-based studies have identified key audio features contributing to SQA, leading to notable advancements in the field. However, they all overlooked the relationship between speech quality and audio distortions. Synthesized speech produced by different VC/TTS systems contains various types of audio distortions, including frequency-dependent noise, temporal and spectral distortions, and dynamic range variations. While some signal processing-based methods utilized some types of audio distortions [32], [33], such as frequency and phase distortion, no attempt was made to integrate audio distortion information into DL models. Moreover, no prior work has integrated audio distortion information from the listeners subjective perspective, thereby aligning the predicted MOS scores more closely with subjective listener perception. On the one hand, audio distortions, both in type and degree, significantly affect listeners speech quality judgments. On the other hand, listeners with varying ages, hearing sensitivities, and levels perceive these distortions differently [34]. Therefore, understanding how distortions are perceived is crucial for accurate audio rating. Further research is needed to extract the perceptual characteristics of audio distortions from synthesized speech, improving the performance of SQA models.

To address existing research gaps, this study explores the potential guidance of audio distortions for deep learning (DL)-based SQA models and introduces a novel feature, the Audio Distortion Token (ADT). ADT captures listeners' perceptions of audio distortions contained in synthesized speech. Leveraging this feature, we propose a novel SQA framework named Audio Distortion Token-Guided Deep MOS Predictor (ADT-MOS). This work represents the first integration of perceptual audio distortion characteristics with DL models, enhancing the correlation between predicted and human-rated MOS values.

To this end, our paper investigates this practical problem, and the contributions are summarized as follows:

1) We propose a novel no-reference deep learning-based SQA framework, **ADTMOS**, that leverages listener-specific perceptual information of audio distortions to enhance the accuracy of MOS predictions. ADTMOS integrates various audio features, including acoustic and metadata information, to capture listeners' subjective perceptions of audio distortions, thereby improving both the predictive accuracy and generalization capability of the SQA model.
2) We propose a low-dimensional, compact embedding, called **audio distortion token**, that reflects listeners subjective perceptions of audio distortions. This distortion-related embedding acts as a bias score for the original MOS at the utterance level, enabling the model output to better align with human ratings.
3) We introduce a novel data augmentation scheme, **Identically-distributed (ID)** data augmentation, designed to address challenges of limited data and imbalanced labels in SQA datasets. This method expands data samples by applying audio augmentations such as speed, volume, and length adjustments, while preserving speech quality. To ensure uniform distribution, we regulate the proportion of augmented samples for each MOS score and VC/TTS system. Experimental results show that this augmentation reduces overfitting and enhances model generalization.
4) We further investigate the impact of different dimensionality unification methods on SQA model performance. Previous work [24] suggests using repetitive padding, which unifies audio embeddings by repeating segments, rather than zero padding, which simply adds zeros. Our experimental and theoretical analysis shows that, compared to repetitive padding, zero padding is more effective and suitable for dimensionality unification.

II. Related Work

Automatic SQA has been an ongoing challenge, with notable advancements over the past decade. SQA methods are typically categorized into full-reference and no-reference approaches, depending on whether a reference speech is required during evaluation. Full-reference SQA relies on a clean reference speech to predict mean opinion scores (MOS), while no-reference SQA estimates MOS without any reference. The following section reviews the related work on these two types of automatic SQA methods.

A. Full-reference SQA

Full-reference SQA methods use the known (reference) speech to predict the quality score for the given test speech. Early methods used time and frequency signal-to-noise ratio (SNR) [8] to evaluate speech quality from speech enhancement or coding systems. These methods require alignment and phase correction between the reference and test speech. Later, several methods [9], [10] utilized the spectral distance or the similarity in the frequency domain between the reference speech and the test speech to perform SQA. Subsequently, several methods [11], [35], [36] have been proposed based on auditory perception models. These methods utilize psychoacoustic knowledge on how humans process tones and noise bands, yielding better results [37]. The PESQ (Perceptual Evaluation of Speech Quality) method [12] first utilizes time alignment and auditory transformation to extract audio distortion characteristics. As an extension of PESQ, POLQA [13] aligns the reference and test speech in time, applies a frequency weighting function, and computes the MOS score.

Although full-reference SQA methods reduce the need for subjective listening tests, they have several limitations. Firstly, the effectiveness of full-reference assessment methods is often confined to specific speech applications and tends to diminish with the emergence of new and varied scenarios. Secondly, they require a clean reference speech, which is unavailable in many applications like VC and TTS. Thirdly, the full-reference SQA usually compares known types of audio distortions and may not be as effective for unknown or non-standard distortions. To address these issues, many no-reference methods have been proposed over the past two decades.

B. No-reference SQA

Unlike full-reference methods, no-reference SQA methods directly assess speech quality without requiring a reference, making them more versatile and applicable in a wide range of scenarios. The research on no-reference SQA can be broadly categorized into two stages. The first stage focused on using various audio features and extracting physical characteristics from the test speech. For instance, Gray et al. [14] applied vocal tract models to identify audio distortions. They first extracted vocal tract shape parameters (e.g., area functions, cavity size) from speech. Then, they analyzed these parameters for physical production violations to enable the detection of potential audio distortions. The E-model [38] predicts MOS values based on audio indicators, including the underlying audio ratio, the transmission delay damage, and the device damage coefficient.

In the second stage, with the rise of deep learning (DL) techniques, DL-based SQA methods have gained increasing attention. These methods can be categorized into three groups. The first group focuses on using feed-forward networks (FFN) or recurrent neural networks (RNN) to extract embeddings correlated with the MOS score from speech's temporal features. Yoshimura et al. [19] proposed a hierarchical framework that integrates the convolution neural network (CNN) and the linear regression (LR) model to predict both the system-level score and the stimulus-level score (i.e., the average of all listeners' ratings for an utterance or several utterances). The predicted system-level score is used as a feature for stimulus-level prediction. Quality-Net [20] predicts the quality of synthesized speech at the utterance level using a Bidirectional Long Short-Term Memory (BLSTM) model, with the BLSTM memory feature improving prediction performance. Later, Lo et al. proposed MOSNet [21], which combines the BLTSM structure of Quality-Net with CNNs to capture temporal and frequency features more efficiently. Yu et al. [39] proposed MetricNet, which leverages label distribution learning and joint speech reconstruction learning to achieve improved prediction accuracy. Although the models mentioned above have achieved comparative performances, they can still be enhanced because the amplitude spectrum they used may not contain enough auditory perception characteristics, such as frequency selectivity and temporal resolution.

Methods in the second category incorporate various cross-domain features, including metadata information about the dataset (e.g., rater groups, system identifiers), various prosodic and linguistic features of the audio, etc. Leng et al. [24] proposed MBNet, which considers all ratings from different listeners for each sample in the dataset. MBNet incorporates the listener ID feature to bring in the subjective preferences of different listeners. Subsequent studies have explored listener-dependent (LD) modeling, focusing on how listener preferences influence the predictive accuracy of SQA. For example, Huang et al. [25] introduced mean listener inference and full listener inference modes and demonstrated the effectiveness of incorporating the listener ID features in SQA. Chinen et al. [26] extended previous work by incorporating metadata features such as listener IDs and system IDs into the model. MOSANet [40] incorporates cross-domain features, including spectral, time-domain, and SSL model outputs, into the model for comprehensive evaluations of speech quality, intelligibility, and audio distortions. However, these methods may overlook certain audio distortions, such as phase distortion, echo, or reverberation, which are not captured by the magnitude spectrum. The loss of certain distortion information might prevent the model from perceiving the speech quality effectively for specific audios, thereby reducing the overall prediction accuracy. Thus, further research is needed to extract various types of audio distortion information and integrate them into DL model training.

The third category focuses on leveraging SSL models to enhance the predictive performance of SQA models. Some models use different large pre-trained SSL models such as Wav2Vec2 [41], HuBert [42], and XLS-R [43] instead of using traditional audio features. Several studies [27], [30] have demonstrated that fine-tuning SSL models can improve the automatic SQA models' generalization ability. Also based on fine-tuned SSL models, some methods improve the prediction model by combining traditional machine learning methods with DL-based strong learners to construct ensemble classifiers. UTMOS [31] adopts a stacked ensemble learning framework that utilizes contrastive learning and phoneme encoding to predict MOS scores. These approaches take advantage of SSL models to enhance the SQA models generalization capability.

The aforementioned methods have explored different DL-based frameworks to predict MOS scores conforming to listeners' ratings. Listeners' ratings are closely related to the speech quality, which is affected by the audio distortions [44]. That is to say, listeners' ratings are influenced by the *listeners' perception of different types of audio distortions*. To ensure that the predicted MOS values closely match the listeners' ratings, SQA methods must account for the perceptual information of audio distortions. However, the extraction of features that are strongly correlated with the perception of audio distortions has not been fully explored.

III. Methodology

A. Overview

The overall architecture of the proposed ADTMOS model has been shown in Fig. 1. ADTMOS consists of three main components: the **feature extraction module**, the **frame-wise MOS score generator**, and the **audio distortion token extractor**. The feature extraction module is responsible for extracting embeddings from the raw audio, which are inputs of the remaining two modules. The frame-wise MOS score generator is responsible for calculating the MOS score for each audio frame, referred to as the frame-wise score embeddings. The frame-wise score embeddings are then globally averaged to obtain the utterance score. The audio distortion token extractor is responsible for extracting ADTs that can reflect the perceptual information of audio distortions. ADTs are then processed into a quantitative score of the degree of audio distortions, named the audio distortion score. The details of these modules are discussed below.

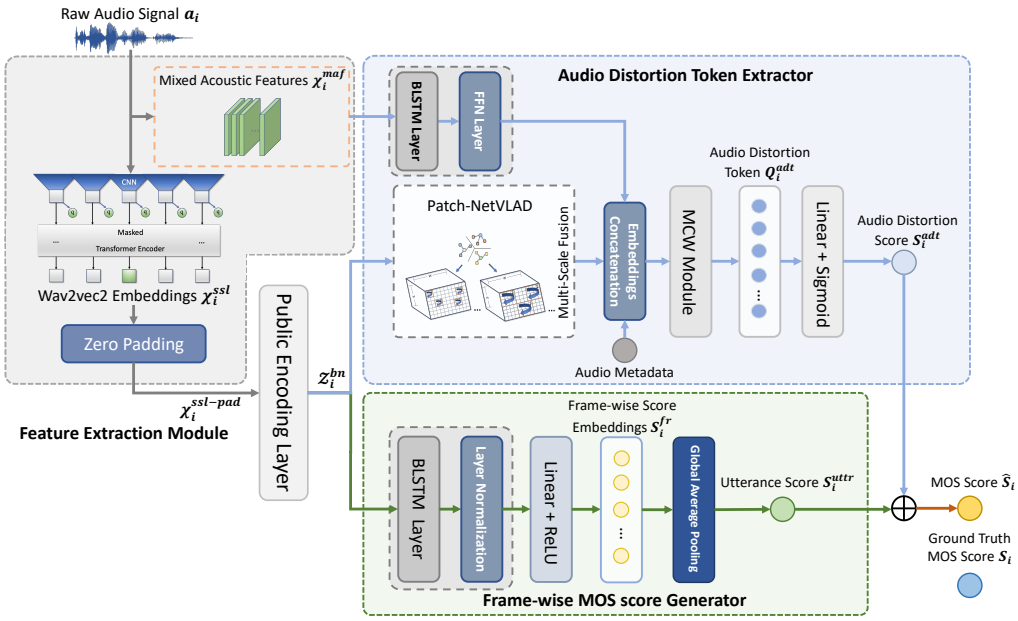


Fig. 1. The overall architecture of ADTMOS. Each raw audio a_i is processed to obtain mixed acoustic features $\mathcal{X}_i^{maf}$ and Wav2Vec2-based embedding $\mathcal{X}_i^{ssl}$. $\mathcal{X}_i^{ssl}$ is then zero-padded and fed into a public encoding layer to get the multichannel audio features $\mathcal{Z}_i^{bn}$. $\mathcal{Z}_i^{bn}$ is fed into the Audio Distortion Token Extractor along with the mixed acoustic features $\mathcal{X}_i^{maf}$ and audio metadata to compute the audio distortion token $\boldsymbol{Q}_i^{adt}$. $\boldsymbol{Q}_i^{adt}$ is then processed with a linear layer to get the audio distortion score S_i^{adt}, which is a quantitative score for the perceptual information of audio distortions. Simultaneously, $\mathcal{Z}_i^{bn}$ is fed into a BLSTM layer and a linear layer to obtain the frame-wise MOS score $\boldsymbol{S}_i^{fr}$, which is then globally averaged to obtain the utterance score S_i^{uttr}.S_i^{adt} and S_i^{uttr} are linearly added to get the predicted MOS score $\hat{S}_i$. The predicted MOS score of each audio is referred to as the utterance-level MOS score.

B. Feature Extraction Module

The feature extraction module extracts audio embeddings from the pre-trained Wav2Vec2 model and various audio features related to audio distortions. Given a dataset $\mathcal{D}$ which contains N audios $a_1, \cdots, a_N$. Firstly, each audio $a_i (i = 1, 2, ..., N)$ is fed into a pre-trained Wav2Vec2 model to get the Wav2Vec2 embeddings $\mathcal{X}_i^{ssl}$. Simultaneously, the mixed acoustic features $\mathcal{X}_i^{maf}$ are extracted from the audio a_i.

1) SSL-based audio features: Tseng et al. [27] showed that using embeddings encoded by SSL models, especially Wav2Vec2, can boost the models' performance for the SQA task. Therefore, we extract SSL-based audio embeddings using the audio representations obtained from Wav2Vec2. To compute SSL-based audio features, the audio a_i is firstly mapped onto the latent representation $\boldsymbol{\mathcal{Z}}_i = [\boldsymbol{z}_{i,1}, \boldsymbol{z}_{i,2}, ..., \boldsymbol{z}_{i,t}]$ using a multi-layer convolutional feature encoder $f : \boldsymbol{a}_i \mapsto \boldsymbol{\mathcal{Z}}_i$, where t represents the audio length. The latent representation $\boldsymbol{\mathcal{Z}}_i$ is fed into a Transformer layer to construct context representations. Following the approach outlined in [45], we use the 7-convolutional layer Wav2Vec2 as the feature extractor. The output from the last convolutional layer of Wav2Vec2 serves as the extracted SSL-based features, which are denoted as $\mathcal{X}_i^{ssl}$.

2) Mixed acoustic features: Speech quality is closely correlated with multiple audio distortion characteristics, including clarity, prosodic naturalness, timbre naturalness, speaking rate, loudness, and prosodic consistency [46]. Different acoustic features can reveal each of these characteristics. For instance, Zero Crossing Rate (ZCR) is closely related to audio clarity. Energy entropy reflects the prosodic naturalness and timbre naturalness to some extent. Therefore, we select the following eight acoustic features to capture the audio distortion characteristics.

- *Zero Crossing Rate (ZCR).* Zero Crossing Rate (ZCR) refers to the number of times the waveform crosses the zero axis within a given time frame. It effectively captures high-frequency components and transient characteristics of the signal. In non-speech-active regions, high-frequency noise artifacts can cause abnormal increases in ZCR, making it a direct indicator of high-frequency reconstruction deficiencies in the synthesis system. Consequently, ZCR serves as a valuable feature for capturing high-frequency distortion in audio.
- *Energy.* Energy represents the average power of an audio per unit of time. Specifically, it is the integration of the squared amplitude of an audio divided by the total length of the audio. Since the intensity and loudness of synthesized speech directly affect the listener's subjective perception, the magnitude of energy can be used as a type of audio distortion feature.
- *Entropy of Energy.* Entropy of Energy quantifies the distribution of energy within an audio signal. To extract this feature, the audio is divided into short-time windows, the energy for each window is calculated, and then the entropy of the energy distribution is computed. In the synthesized speech, as the quality deteriorates, the energy entropy increases, reflecting greater irregularity and complexity in the energy distribution.
- *Spectral Centroid.* Spectral centroid is a measure that quantifies the distribution of high and low-frequency components in audio and can reflect the clarity and loudness of synthesized speech. The audio will sound sharp and piercing if the spectral centroid is too high.

If it is too low, the audio will sound dull and muffled.

- *Spectral Spread*. Spectral spread is used to measure the width and the shape of the spectral distribution, which can reflect the sound quality of synthetic speech. The speech will sound monotonous and lack expressiveness if the spectral spread is too small. If it is too large, the speech will sound chaotic. Therefore, this feature reflects the distortion of synthetic speech from the frequency range aspect.
- *Spectral Entropy*. Spectral entropy quantifies the evenness of the energy distribution in the frequency spectrum of the audio. A higher spectral entropy value indicates a more even spectral energy distribution, which suggests that the audio is more clear and more natural-sounding.
- *Spectral Flux*. Spectral flux measures the rate of the change of the spectral content between adjacent frames in the audio. It is used to quantify the smoothness and variability of the audio. The magnitude of spectral flux can reflect the synthesized speech's perceived naturalness or sharpness.
- *Spectral Rolloff*. Spectral Rolloff is the frequency below which a certain percentage of the total energy of the signal is contained. Therefore, Spectral Rolloff reflects the audio's spectral characteristics and energy distribution. The energy of the human voice is usually concentrated in the lower frequency range so that the Spectral Rolloff can reflect the high-frequency noise of the human voice.

The above 8 features are concatenated together to construct a set of mixed acoustic features $\chi_i^{maf} \in \mathbb{R}^{t \times d}$, where t represents the audio length and d represents the dimension of the mixed acoustic features.

3) Dimensionality Unification: Due to different lengths of a_i, the dimension of the extracted features might be inconsistent, which is inconvenient for training. Therefore, it is necessary to unify the dimensions of the input audio features. The two commonly used methods of dimensionality unification for SQA tasks are repetitive padding [24] and zero padding [21]. Repetitive padding duplicates a short audio segment and concatenates the segment with the original one. By contrast, zero padding only adds zero values to the end of the short audio. Leng et al. [24] claim that repetitive padding does not change speech quality and that using it has better results in their experiments. Accordingly, most recent work [24]–[26], [28], [31] applied the repetitive padding operation for dimensionality unification. However, such a claim is not suspicious because repetitive padding changes the speech content. By contrast, the zero padding approach adds some silence to the audio, which will not affect speech quality. Theoretically, zero padding is more suitable as a preprocessing step. However, which operation is more appropriate for the speech quality evaluation still needs further exploration. To this end, we conducted extended experiments on their impacts on the model performances. The experimental results shown in the supplemental material indicate that the model integrating zero padding performs better in the experiment. Interestingly, Leng et al. [24] claimed that repetitive padding allows the audio features to have a more stable mean and variance when trained using batch normalization. However, we did not observe that repetitive padding provides such a benefit from our experiments. By contrast, this method changes the speech quality of each audio, which may negatively impact the model performance. Therefore, zero padding is adopted in our final solution to perform dimensionality unification. Specifically, we apply zero padding to each acoustic feature $\mathcal{X}_i^{maf}$, and then concatenate them into an $M \times d$ feature vector, where M refers to the maximum length of each acoustic feature. We pad the SSL-based embeddings $\mathcal{X}_i^{ssl}$ with zeros to match the feature size of the audio with the maximum length M in the dataset $\mathcal{D}$ and denote the padded embedding as $\mathcal{X}_i^{ssl-pad}$.

C. Public Encoding Layer

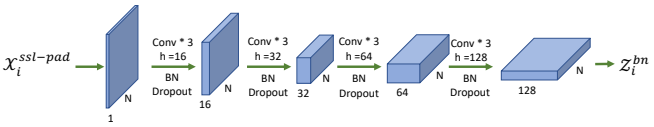


Fig. 2. The detailed architecture of the public encoding layer. *Conv* stands for 2D convolutional layers, *h* stands for the channel size, *BN* stands for batch normalization, and *Dropout* means leveraging a dropout layer after every 3 convolutional layers.

As shown in Fig. 2, the public encoding layer comprises four convolutional layers. Each layer consists of 3 2D convolutional layers, followed by dropout and batch normalization (BN). Similar to MBNet [24], we incorporate dropout and batch normalization operations in our model to mitigate overfitting and gradient explosion compared with the CNN module of MOSNet [21]. The output of the public encoding layer is denoted as $\mathcal{Z}_i^{bn}$.

D. Frame-wise MOS Score Generator

The frame-wise MOS score generator processes the output of the public encoding layer $\mathcal{Z}_i^{bn}$ to compute the frame-wise MOS score and the utterance score, where the frame-wise MOS score is a vector representing the MOS prediction scores for each frame of the audio. Similar to MOSNet [21], $\mathcal{Z}_i^{bn}$ is first passed through a BLSTM layer, followed by a linear layer with ReLU activation to derive the frame-wise MOS score $\boldsymbol{S}_i^{fr}$ (as shown in Fig. 1). Then, the frame-wise MOS score is passed through a global average pooling operation to obtain the utterance score S_i^{uttr}. A layer normalization (LN) [47] operation is incorporated after the BLSTM layer to further prevent gradient vanishing and speed up model convergence.

E. Audio Distortion Token Extractor

The audio distortion token extractor is designed to extract the perceptual information of audio distortions. Wilson et al. [44] demonstrated that speech quality is closely related to different types of audio distortions. Speech synthesized by different VC/TTS systems exhibits distinct audio distortion characteristics, which give listeners different impressions of the quality of the speech. Therefore, this module is designed to leverage different audio distortions and extract the embeddings regarding the listeners' perceptions of audio distortions.

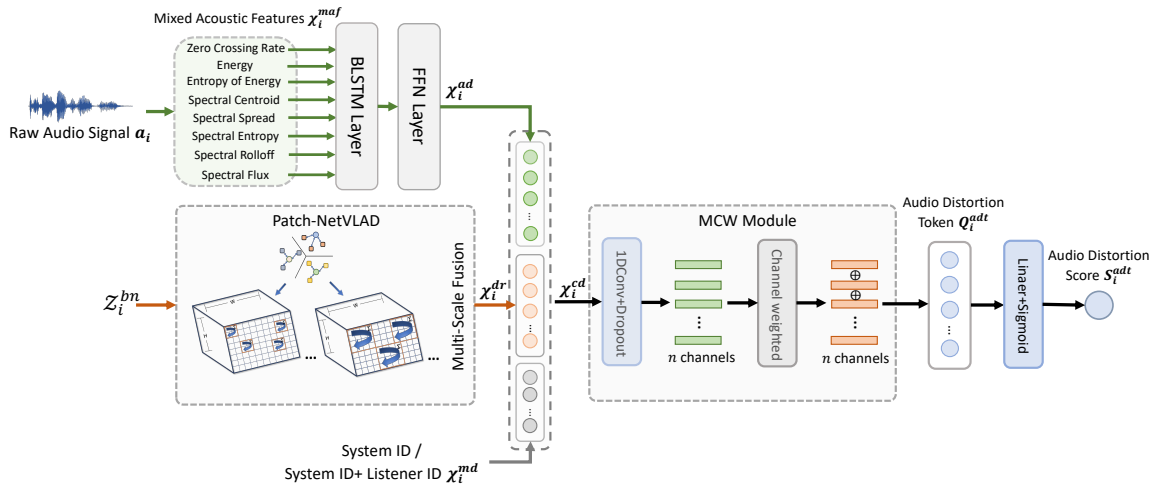


Fig. 3. The architecture of the Audio Distortion Token Extractor. It feeds the output of the public encoding layer to the Patch-NetVLAD module to obtain a relatively compact representation of the audio features $\boldsymbol{\chi}_i^{dr}$. $\boldsymbol{\chi}_i^{dr}$ are concatenated with the audio metadata features (the system ID and the listener ID), as well as the audio distortion embeddings $\boldsymbol{\chi}_i^{ad}$ that have been processed by a BLSTM layer and an FFN layer, to form a cross-domain feature $\boldsymbol{\chi}_i^{cd}$. The cross-domain features $\boldsymbol{\chi}_i^{cd}$ are then fed into an MCW layer to extract the perceptual information embedding for audio distortion features, named the audio distortion token $\boldsymbol{Q}_i^{adt}$. The audio distortion token $\boldsymbol{Q}_i^{adt}$ is finally fed into a linear layer to fetch the ADT score S_i^{adt}.

The structure of the audio distortion token extractor is shown in Fig. 3. It mainly consists of the **Patch-NetVLAD Module**, the **Embeddings Concatenation Module**, and the **MCW** Module. The audio distortion token extractor uses the convolution process and a channel-weighting operation to extract the listeners' perceptual information of audio distortions to obtain the audio distortion score, which can help the model output more accurate MOS scores. To our knowledge, we are the first to use DL techniques to extract information about the listener's perceptions of audio distortions.

1) Patch-NetVLAD Module: This module utilizes Patch-NetVLAD [48] to perform clustering and dimensionality reduction on audio features. It reduces the dimension of the input features through learnable parameter pooling. Patch-NetVLAD was initially used for scene recognition in computer vision and is an optimization of NetVLAD [49]. It improves the recall rate of scene recognition by accelerating the computation of multi-scale patch feature descriptors. Tang et al. [50] leverage NetVLAD to simultaneously encode video and audio features for video classification, which demonstrates the efficacy of NetVLAD in processing audio features as well. SQA is a typical audio-related classification task, so Patch-NetVLAD is suitable for aggregating distinctive audio features to obtain the compact encoding of audio features in this task. Specifically, the audio features $\mathcal{Z}_i^{bn}$ extracted by the public coding layer in Section III-C are first segmented into localized patches at different scales, where the jth patching result under window size s is calculated as:

$$\boldsymbol{\mathcal{P}}_s = \{\boldsymbol{P}_{s,j}\}_{j=1}^{N_s}, \quad \boldsymbol{P}_{s,j} \in \mathbb{R}^{s \times D} \tag{1}$$

where N_s is the number of patches and D is the number of channels in $\mathcal{Z}_i^{bn}$. Then each trunk $\boldsymbol{\mathcal{P}}_{s,j} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_s]^\top$ undergoes NetVLAD-based feature aggregation, which can be expressed as:

$$\boldsymbol{V}_j(k) = \sum_{t=1}^{s} \alpha_k(\boldsymbol{x}_t)(\boldsymbol{x}_t - \boldsymbol{c}_k), \tag{2}$$

$$\alpha_k(\boldsymbol{x}_t) = \frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x}_t + b_k)}{\sum_{k'=1}^{K} \exp(\boldsymbol{w}_{k'}^\top x_t + b_{k'})}, \tag{3}$$

where $c_k \in \mathbb{R}^D$ denotes learnable cluster center k , $\mathbf{w}_k \in \mathbb{R}^D, b_k \in \mathbb{R}$ denotes soft-assignment parameters, $V \in \mathbb{R}^{k \times D}$ denotes output descriptor for the patch. Then, the features from different scales are fused hierarchically:

$$\boldsymbol{V}_s = \frac{1}{N_s} \sum_{j=1}^{N_s} \boldsymbol{V}_{s,j} \quad \in \mathbb{R}^{k \times D}, \tag{4}$$

$$\boldsymbol{V}_{concat} = [\boldsymbol{V}_{s_1}; \boldsymbol{V}_{s_2}; ...; \boldsymbol{V}_{s_m}] \quad \in \mathbb{R}^{(m \cdot k) \times D}, \tag{5}$$

where the hyperparameter m denotes the number of scales. Finally, a projection layer is utilized to generate the compact encoding $\boldsymbol{\chi}_i^{dr}$:

$$\boldsymbol{\chi}_i^{dr} = Reshape\left(\boldsymbol{W}_{proj} \cdot \boldsymbol{V}_{\text{concat}} + \boldsymbol{b}_{proj}\right) \in \mathbb{R}^{(m \cdot k) \times c}, \tag{6}$$

where $\boldsymbol{W}_{proj} \in \mathbb{R}^{D \times c}$ reduces each cluster dimension from D to c, k represents the number of cluster centers in Patch-NetVLAD and c is the encoding dimension of Patch-NetVLAD. The parameters of Patch-NetVLAD will be jointly optimized with those of ADTMOS during training.

2) Embeddings Concatenation: After aggregating distinctive audio features using the Patch-NetVLAD module, we incorporate the metadata features from the audio into the compact representation of low-dimensional audio features $\boldsymbol{\chi}_i^{dr}$. Since the system IDs provide information about the VC/TTS system to which the audio a_i belongs and the listener IDs reflect information about the listener for each rating, they are integrated and used as metadata inputs. Because there is no semantic correlation between the system ID and the listener

ID of the audio, we encode them separately as one-hot vectors. Then, we concatenate these two sets of one-hot vectors as

$$\boldsymbol{\chi}_i^{md} = OneHot(\omega_i) \oplus OneHot(l_i), \tag{7}$$

where ω_i denotes the VC/TTS system corresponding to audio a_i, l_i indicates the listener ID of the rater who evaluated audio a_i. In addition, we can also directly use the encoded system IDs as the audio metadata features without introducing listener preference information.

In practice, both during the rating phase of VC/TTS competitions and the inference of SQA tasks, system metadata (i.e., system and listener IDs) are not always feasible. Therefore, similar to [26], during training, we introduce an **unknown** category to deal with the situation in which metadata information is unavailable. Specifically, we add an **unknown system** for system IDs in the existing system categories in the training set. For listener IDs, we add an **unknown listener** to the existing set of listeners during training. To ensure ADTMOS accurately extracts perceptual information of audio distortions, we incorporate **random masking** into the system metadata during training. It means that, with a given probability p, we randomly mask the system ID of the training sample and replace it with the unknown system.

For the mixed acoustic features χ_i^{maf} of audio a_i obtained in Section III-B2, we use a BLSTM layer and an FFN layer to obtain the implicit features of audio distortions $\boldsymbol{\chi}_i^{ad}$:

$$\boldsymbol{\chi}_i^{ad} = W_i^{maf} BLSTM\left(\chi_i^{maf}\right) + \boldsymbol{b}_i^{maf}, \tag{8}$$

where W_i^{maf} and $\boldsymbol{b}_i^{maf}$ are the learnable weights and biases of the FFN layer, respectively. Finally, we concatenate the aggregated feature vectors $\boldsymbol{\chi}_i^{dr}$, the audio metadata $\boldsymbol{\chi}_i^{md}$ and $\boldsymbol{\chi}_i^{ad}$ to form a cross-domain embedding:

$$\boldsymbol{\chi}_i^{cd} = \boldsymbol{\chi}_i^{dr} \oplus \boldsymbol{\chi}_i^{md} \oplus \boldsymbol{\chi}_i^{ad}. \tag{9}$$

A combination of these features will make ADTMOS better model how listeners perceive audio distortions. The rationale is that, on the one hand, the features of the audio feature $\boldsymbol{\chi}_i^{dr}$ can provide the overall information about the audio. On the other hand, the audio distortion embedding $\boldsymbol{\chi}_i^{ad}$ is an implicit feature related to audio distortions, which can force our model to learn listeners perceptions of certain specific distortion characteristics.

3) MCW Module: The concatenated cross-domain embeddings $\boldsymbol{\chi}_i^{cd}$ are then fed into the MCW module to obtain the audio distortion token related to listeners' perceptions of audio distortions. Many methods in image quality assessment adopted CNNs to extract features of various granularities and types of image distortions [51]. Inspired by these methods, we designed the MCW module to extract listeners' perceptual information of audio distortions at different granularities.

In the MCW module, the cross-domain embeddings $\boldsymbol{\chi}_i^{cd}$ are first fed into a 1D convolutional layer to get multichannel cross-domain features. This process can be expressed as

$$\left[\boldsymbol{Z}_i^1, \boldsymbol{Z}_i^2, ..., \boldsymbol{Z}_i^C\right] = \boldsymbol{\chi}_i^{cd} * k(n, n), \tag{10}$$

where the number of channels C is a hyperparameter indicating the coarseness of the granularity of the listener's perceptions of audio distortions. A small C indicates fewer channels and usually corresponds to coarse-grained perceptions of audio distortions, e.g., the perception of changes in the overall loudness, the dynamic range of the audio, and the high-frequency noise. By contrast, a large C corresponds to fine-grained perceptions, e.g., vocabulary-level discrimination or harmonic distortions.

Then, we fuse the multichannel perception features of audio distortions using a channel-weighted operation. For the cth $(c = 1, 2, ..., C)$ channel vector $\boldsymbol{Z}_i^c$, this process can be expressed as

$$\alpha_i^c = Sigmoid\left(W_i^c Z_i^c + \boldsymbol{b}_i^c\right), \tag{11}$$

where $\boldsymbol{\alpha}_i^c$ is the weight score vector of the cth channel features, W_i^c and $\boldsymbol{b}_i^c$ are the learnable parameters of the channel-weighted operation. We perform a weighted fusion of all channel features by dot product and summation, and the process can be expressed as

$$\boldsymbol{Q}_i^{adt} = ReLU\left(W_i^{adt} \sum_{c=1}^{C} < \alpha_i^c \boldsymbol{Z}_i^c > + \boldsymbol{b}_i^{adt}\right), \tag{12}$$

where $\boldsymbol{Q}_i^{adt}$ represents the audio distortion token of a_i, $<>$ denotes the dot product operation, W_i^{adt} and $\boldsymbol{b}_i^{adt}$ are the learnable parameters associated with the weighted fusion operation. The audio distortion token $\boldsymbol{Q}_i^{adt}$ is a low-dimensional and compact vector representation regarding the perceptual information of audio distortions. The final audio distortion score S_i^{adt} can be calculated through a fully connected layer:

$$S_i^{adt} = Sigmoid\left(\boldsymbol{W}_i^{adt} \boldsymbol{Q}_i^{adt} + b_i^{adt}\right), \tag{13}$$

where S_i^{adt} denotes the audio distortion score of audio a_i, $\boldsymbol{W}_i^{adt}$ and b_i^{adt} are learnable parameters.

F. Calculation of the Predicted MOS score

After obtaining the utterance score S_i^{uttr} and the audio distortion score S_i^{adt}, the predicted MOS score $\hat{S}_i$ can be computed via a linear weighting operation:

$$\hat{S}_i = S_i^{uttr} + \alpha S_i^{adt}, \tag{14}$$

where S_i denotes the predicted MOS score of a_i. α is a scaling factor that determines the relative importance of the audio distortion score in the predicted MOS score.

G. Loss Function

The loss function of ADTMOS consists of two loss terms: the loss of the predicted MOS scores of each frame and the loss of overall MOS score prediction. Previous work [22], [24]–[26] often used the mean squared error (MSE) loss or the clipped MSE loss, but outliers can easily dominate the direction of gradient updates. Therefore, in this work, Smooth L1 Loss [52] is used as the loss function to increase the convergence smoothness as shown in Eq. (15).

$$\mathcal{L}_\delta(\hat{y}, y) = \begin{cases} \frac{1}{2}(\hat{y} - y)^2, & |y - \hat{y}| \leqslant \delta \\ \delta \cdot \left(|\hat{y} - y| - \frac{1}{2}\delta\right), & |y - \hat{y}| > \delta. \end{cases} \tag{15}$$

In Eq. (15), the hyperparameter δ is the smoothing parameter in the Smooth L1 Loss. The loss function for ADTMOS is calculated as

$$\mathcal{L}_{mos} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{L}_{\delta}\left(\hat{S}_i, S_i\right) + \frac{\beta}{F_i}\sum_{f=1}^{F_i}\mathcal{L}_{\delta}\left(S_i^{fr}[f] + \alpha S_i^{adt}, S_i[f]\right)\right), \tag{16}$$

where $S_i[f]$ denotes the ground truth MOS score of the fth frame, and $S_i^{fr}[f]$ denotes the predicted MOS score of fth frame. β is a hyperparameter that balances two parts of the loss function. To calculate the frame-wise score loss, we need to expand the actual MOS values into a vector according to the number of frames in the audio for comparison with predicted frame-wise scores. Since the predicted frame-wise scores $\boldsymbol{S}_i^{fr}$ also need to incorporate bias from audio perceptual distortion prediction to obtain final MOS predictions, this bias term, the audio distortion score S_i^{adt}, needs to be added, and adjusted by the specific weighting factor α mentioned in Section III-G.

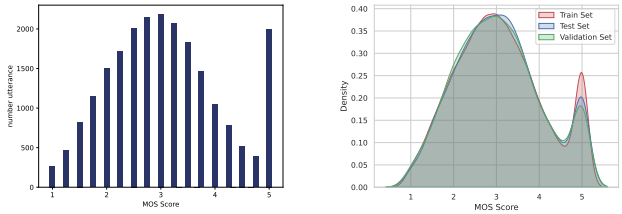
IV. Experiment and Analysis

A. Dataset

To evaluate the predictive performance and generalization ability of ADTMOS, we choose the expanded Voice Conversion Challenge 2018 (VCC2018) [53] dataset, i.e., the VCC2018-CSMSC dataset, and the BVCC [54] dataset. Compared with the BVCC dataset, VCC2018-CSMSC contains fewer VC/TTS systems but more utterances per system, which can test the model's prediction accuracy on a larger scale. On the contrary, BVCC contains more VC/TTS systems with fewer utterances. Its test set includes unseen systems, speakers, and listeners, which can test the model's generalization ability in dealing with unknown systems.

1) VCC2018-CSMSC Dataset: The original VCC2018 dataset contains 20,580 audio samples submitted by 38 VC/TTS systems. A total of 267 experts participated in the rating task of VCC2018, and the quality of each speech sample was rated by four experts.

Most of the speech samples' MOS values in the VCC2018 dataset are 2-4, so there were remarkably fewer high-scored and low-scored speech samples. Compared with low-quality speech samples, high-quality speech samples, such as TV series or news broadcasts recorded in the studio, are comparatively easier to obtain. Following this idea, we retrieved TV series samples from the Chinese Standard Mandarin Speech corpus (CSMSC) [55] and added them to the VCC2018 dataset. CSMSC is an open dataset containing 12 hours of professional standard Mandarin female voice recordings with a signal-to-noise ratio of no less than 35 dB in the recording environment. Considering the high quality of the speeches in the CSMSC dataset, their MOS scores are set to 5. In our experiments, 1,780 clear speeches from the CSMSC dataset with MOS scores 5 were added to the VCC2018 dataset to increase the proportion of high-quality speech samples. As a comparison, adding low-scored speech samples is trickier and requires additional ratings. However, the new rating distribution may be inconsistent with the original VCC2018 dataset's rating distribution. Therefore, we did not add new low-quality speech samples. The details of the expanded VCC2018 dataset, namely, VCC2018-CSMSC, are shown in Table I. We assume that one additional virtual listener rates all speech samples from CSMSC, and all samples from CSMSC are considered to be generated by a new VC/TTS system.



(a) (b)

Fig. 4. (a) The MOS score distribution of the VCC2018-CSMSC dataset. The horizontal axis represents the MOS values, while the vertical axis represents the number of audio samples corresponding to each MOS value; (b) KDE diagram of the train, valid, and test set using the hold-out splitting method. The horizontal axis represents the MOS values, while the vertical axis represents the density corresponding to each MOS value.

Fig. 4(a) shows the MOS score distribution in the VCC2018-CSMSC dataset. Since the overall distribution of the training set is similar to a normal distribution, a simple hold-out method can ensure that the training set, validation set, and test set virtually have the same distribution. We used the hold-out method to divide the dataset into the train, valid, and test sets 1000 times, and employ the Wasserstein distance [30] in each iteration to assess the distribution differences. Finally, we pick the dataset division with the smallest Wasserstein distance as the final dataset division. As shown in Fig. 4(b), we use Kernel Density Estimation (KDE) [56] to show the distributions of the training set, validation set, and test set after using the hold-out method to split the dataset.

2) BVCC dataset: The BVCC dataset was provided in the main track of Voice MOS Challenge 2022 [54]. It includes 7,106 audio clips with a total duration of 8.02 hours. The duration of each audio ranges from 0.82 seconds to 45.74 seconds. The time duration varies significantly because the BVCC dataset contains mixed data from multiple datasets including Bizzard Challenge 2008 - 2011, and VCC2016, VCC2020 with varying audio lengths.

As shown in Table I, the audio samples in the BVCC dataset come from 187 VC/TTS systems, 175 of which have appeared in the training set. The validation set contains audio samples from 6 new systems in addition to the existing systems, and the test set contains audio samples from 12 new systems in addition to the existing systems, respectively. These unknown systems will be used to test the performance of models when encountering unknown VC/TTS systems, which can reflect the generation abilities of these models.

B. Data Augmentation

The proportions of low-scored and high-scored audio samples in the two datasets above are relatively small. It often

TABLE I
DETAILS OF THE VCC2018-CSMSC DATASETS AND THE BVCC DATASETS.

	VCC2018-CSMSC	BVCC		
	-	train	valid	test
# of ratings	82,720 (80,940+1,780)	39,792	8,528	8,528
# of audios	22,360	4,974	1,066	1,066
# of audios per VC/TTS system	560-1,780 (avg: 828.1)	12-36 (avg: 29.4)	1-37 (avg: 5.9)	1-38 (avg: 5.7)
# of VC/TTS systems	27	175	175 (seen)+6 (unseen)	175 (seen)+12 (unseen)
# of listeners	225	288	288 (seen) + 8 (unseen)	288 (seen)+16 (unseen)

results in training bias of the model and hampers the model's performance. To address this issue, we propose two data augmentation methods, namely Identically Distributed (**ID**) Data Augmentation and Proportion-Aware (**PA**) Data Augmentation, the details of which are given below.

1) Preprocessing: In prior investigations [30], speed change and silence addition were commonly used as the preprocessing step before data augmentation. Denote the original audio as a. Three preprocessing methods are applied to a, including clipping the audio, adjusting the volume, and altering the playback speed.

- **Clip the Audio**. If the duration of audio a is longer than a threshold, it will be clipped to generate a short audio a_{clip}:

$$f_{clip}(a, t_0, t_d) = a[t_0, t_0 + t_d], \\ 0 \leqslant t_0 \leqslant \frac{n_0}{fs}, 0 < t_d \leqslant \frac{n - n_0}{fs}. \tag{17}$$

  Here, t_0 and t_d represent the starting time and duration, respectively. fs is the sampling rate of the original audio.
- **Change Audio's Volume**. We increase the volume of the audio by amplifying the original volume with the factor of l:

$$g_{cv}(a, l) = min(max(l{*}a[t], -1), 1), 0 \leqslant t \leqslant \frac{n}{fs}, l > 0. \tag{18}$$

  We use a min-max operation to ensure that the sample values remain within the range of (-1, 1) after changing the loudness.
- **Change Audio's Speed**. Perform time-stretching on the original audio using the Lagrange interpolation method. In this study, the speed change factor v is between 0.95 and 1.05.

We applied three data preprocessing methods on each audio a_i:

$$a_{aug} = h_{cs}(g_{cv}(f_{clip}(a_i, t_0, t_d,), l), v) \tag{19}$$

Huang et al. [54] mentioned that, for SQA tasks, the training set's distribution could significantly impact the model performance. On the one hand, the uniform distribution of samples with different MOS scores ensures that the model is adequately trained for each MOS score. On the other hand, the identical distributions between the training and testing sets can avoid the impact of distributional changes on the model's prediction performance. To investigate the impact of different training data distributions on the model's performance, we introduce two data augmentation methods that produce two types mentioned above of data distribution.

2) Identically Distributed Data Augmentation: The ID method ensures the distribution consistency between the augmented and original training sets. Expressly, for a given SQA training set $\mathcal{D}_{train}$, assume we have N_{aug} as the number of samples designated for data augmentation. For each MOS score S in the training set, we need to randomly sample $N_{aug} * \frac{|\mathcal{D}^S_{train}|}{|\mathcal{D}_{train}|}$ audios with a replacement for data augmentation, where $|\mathcal{D}^S_{train}|$ denotes the number of audios in training set with an MOS score of S. After applying the augmentation process shown in Eq. (19) to all the randomly sampled audios respectively (note that one audio may be selected and augmented multiple times), we can obtain the desired augmented set $\mathcal{D}_{aug}$. The ID data augmentation can ensure that the distribution of the augmented training set is consistent with the overall distribution.

3) Proportion-Aware Data Augmentation: Due to the high model accuracy of recent VC/TTS systems, the SQA datasets generally suffer from a lack of low-MOS and high-MOS data, which leads to poor performance of recent SQA models in predicting extreme MOS values. To address this issue, the PA method is proposed to balance the label distribution in the dataset. Assuming that in a given SQA training set $\mathcal{D}_{train}$, $N^{'}_{aug}$ represents the number of audio samples per MOS score after data augmentation. Compared to the ID data augmentation method, the only difference is that the number of random sampling for each MOS score S is $N^{'}_{aug} - |\mathcal{D}^S_{train}|$. By utilizing the PA data augmentation, the low-scored and high-scored data sets are effectively expanded, making the distribution of training samples more balanced, thereby further improving the accuracy of MOS prediction.

C. Implement Details

1) Experimental Settings: For the Wav2Vec2 model, both the chinese-wav2vec2-large[1] and the wav2vec2-base-960h[2] model are used to process raw audios in Chinese and English. The former model was pre-trained and fine-tuned on 10,000 hours of audio of the WenetSpeech L subset [57], and the latter was pre-trained and fine-tuned on 960 hours of audio of the Librispeech dataset [58]. The number of hidden units in BLSTM for all our models is 128. The dimension of fully connected layers for all our models is 128. The dropout rate for the convolutional layers in the public encoding layer is set to 0.3. The channels in the MCW module C are set to 6. The value setting of C is further discussed in the supplemental material.

[1]https://huggingface.co/TencentGameMate/chinese-wav2vec2-large
[2]https://huggingface.co/facebook/wav2vec2-base-960h

For training details, we trained ADTMOS using two NVIDIA RTX3070Ti GPUs. The model was trained with the AdamW optimizer [59] with a weight decay coefficient of 0.0001. Batch size and learning rate are set to 64 and 0.0001, respectively. Early stopping was applied based on the overall loss of the validation set with a patience of 20 epochs.

2) Configuration of baseline models: To demonstrate its effectiveness, ADTMOS was compared with four competitive SQA algorithms, including MOSNet [21], MBNet [24], and two variants of LDNet [25].

- **MOSNet** [21] An end-to-end CNN-BLSTM framework that integrates frame-level MOS score and utterance-level MOS score to predict the final MOS score. In our experiment, MOSNet was implemented using CNN-BLSTM with the configuration given in [21].
- **MBNet** [24] An SQA model that utilizes listener IDs during training to learn the listener preference. This model consists of two sub-networks: the MeanNet, which is trained to predict the mean MOS score, and the BiasNet, which is trained to predict the listener's bias score. The outputs of the two sub-networks are added together to produce the predicted MOS score.
- **LDNet-MN** [25] A variant of LDNet that incorporates a Mean-Bias (MB) architecture just like MBNet. LDNet-MN utilizes a single-layered FFN to predict the mean MOS score and a BLSTM-based RNN decoder to predict the listener's bias score.
- **LDNet-ML** [25] A variant of LDNet that directly incorporates listener IDs and adds a new class (i.e., the mean listener) to predict both the listener's bias score and the mean MOS score. Unlike LDNet-MN, LDNet-ML simply utilizes a single-layered FFN as a decoder to predict MOS scores.

3) Evaluating Metrics: In SQA tasks, the model performance must be measured both at the utterance and system levels. At the utterance level, we compare the predicted MOS value $\hat{S}_i$ of audio a_i with its actual MOS value. At the system level, however, we calculate the average MOS value of all audios for each VC/TTS system as the predicted MOS value for the synthesized audio of that system. The ground truth MOS value of the VC/TTS system is calculated similarly. By comparing the predicted MOS score of individual VC/TTS systems with their ground truth MOS values, we can determine the accuracy of the SQA model in assessing the quality of speech synthesized by various VC/TTS algorithms.

As for the metrics, apart from the commonly used evaluation metrics such as Linear Correlation Coefficient (LCC), Spearman's Rank Correlation Coefficient (SRCC), Mean Square Error (MSE) and Kendall's Tau (KTAU), Mean Absolute Error (MAE) and Coefficient of Determination (R^2) are used as additional evaluation measurements in the experiments. MAE is less sensitive to outliers than MSE and can be used as a supplement to MSE. As a metric used to evaluate the goodness of fit of regression models, R^2 measures the degree to which the model explains the variance in observed data.

In addition, we also define a new metric called **MOS Score Accuracy** (MSA) to measure the accuracy of the model in predicting audio ratings:

$$MSA_\tau = \frac{\sum_{i=1}^{N} \mathbb{N}\left(\left|\hat{S}_i - S_i\right| < \tau\right)}{N} \tag{20}$$

where $\mathbb{N}(\cdot)$ denotes the indicator function whose output is 1 when the condition is true; otherwise, 0. τ represents the threshold for determining whether the predicted MOS of audio is correct or not. In our experiments, we set τ to 1 at the utterance level and 0.5 at the system level.

D. Performance Evaluation

Table II and Table III demonstrate the experimental results of the VCC2018-CSMSC dataset and the BVCC dataset of the eight methods which include four baseline models and four variants of ADTMOS. In the experiment, four ADTMOS variants, namely ADTMOS-SI, ADTMOS-SI-LI, ADTMOS-SI-ID, and ADTMOS-SI-PA, were trained using different audio metadata and augmentation methods. ADTMOS-SI uses system IDs as the audio metadata input $\mathcal{X}_i^{md}$ mentioned in Section III-E2. ADTMOS-SI-LI uses both system IDs and listener IDs (LI) as the audio metadata inputs. ADTMOS-SI-ID and ADTMOS-SI-PA use the ID and PA data augmentation method in ADTMOS-SI, respectively. Data augmentation becomes unnecessary since utilizing listener IDs significantly increases the training data volume. Therefore, we do not validate ADTMOS-SI-LI-ID/PA, which are variants that utilize both listener IDs and ID/PA data augmentation methods.

1) Performance Evaluation on VCC2018-CSMSC dataset: In Table II, compared to MOSNet, all LD-based models, including MBNet, LDNet-MN, and LDNet-ML, achieve much better performances. These results suggest that integrating listener preferences has a promoting effect on the performance of automatic SQA models. Among the two LDNet variants, LDNet-ML performs slightly better than LDNet-MN with its utterance-level LCC of 0.775, SRCC of 0.723, and MSE of 0.453. It indicates that using a single network to predict both the LD score and the mean score is more efficient than using a separate "MeanNet" and "BiasNet". Consequently, our work did not employ a separate "MeanNet" to predict the LD score.

It can be observed that the three ADTMOS variants, i.e., ADTMOS-SI, ADTMOS-SI-LI, and ADTMOS-SI-ID, overwhelm all the baseline models. Specifically, ADTMOS-SI-LI obtains the utterance-level SRCC of **0.730** and the system-level SRCC of **0.965** compared to LDNet-ML of 0.723 and 0.957. This could be attributed to the fact that ADTMOS-SI-LI unifies more audio metadata and perceptual information of audio distortions, enhancing the model's ability to distinguish between audios of similar quality. In addition, ADTMOS-SI-LI shows the most significant improvement in utterance-level MSE with an increase of 12.5% compared to LDNet-ML, reaching **0.396**. It indicates that using the outputs of Wav2Vec2 models as audio embeddings instead of the amplitude spectrum can better capture the features of speech quality. Compared to ADTMOS-SI, the most notable improvement for ADTMOS-SI-LI is the system-level SRCC, which increases from 0.944 to 0.965. This indicates that incorporating system IDs significantly enhances the ability to assess different VC/TTS

TABLE II
THE UTTERANCE-LEVEL AND THE SYSTEM-LEVEL EXPERIMENTAL RESULTS ON VCC2018-CSMSC DATASETS.

Method	utterance-level							system-level						
	LCC↑	SRCC ↑	MSE ↓	KTAU ↑	MAE ↓	R2 ↑	MSA_1↑	LCC ↑	SRCC ↑	MSE ↓	KTAU↑	MAE ↓	R2 ↑	$MSA_{0.5}$↑
MOSNet [21]	0.721	0.622	0.490	0.468	0.553	0.516	83.89%	0.900	0.871	0.115	0.721	0.227	0.775	100%
MBNet [24]	0.791	0.722	0.538	0.558	0.556	0.514	82.29%	0.982	0.949	0.144	0.835	0.356	0.731	100%
LDNet-MN [25]	0.768	0.721	0.459	0.553	0.540	0.585	86.28%	0.982	0.957	0.022	**0.863**	0.111	0.958	100%
LDNet-ML [25]	0.775	0.723	0.453	0.550	0.536	0.591	86.07%	0.983	0.939	0.027	0.835	0.121	0.950	100%
ADTMOS-SI	0.789	0.726	0.407	0.560	0.491	0.614	87.97%	0.986	0.944	**0.016**	0.858	**0.086**	**0.970**	100%
ADTMOS-SI-LI	**0.794**	**0.730**	0.396	**0.565**	0.486	0.625	**88.61**%	0.985	**0.965**	0.017	**0.863**	0.100	0.968	100%
ADTMOS-SI-ID	**0.794**	0.729	**0.393**	0.564	**0.483**	**0.627**	88.57%	**0.987**	0.955	0.021	0.846	0.108	0.961	100%
ADTMOS-SI-PA	0.765	0.700	0.455	0.537	0.516	0.569	85.93%	0.977	0.955	0.027	0.846	0.112	0.950	100%

systems' overall speech quality. Specifically, providing system ID metadata during training enables ADTMOS to grasp the implicit features of speech quality in the synthesized audio from different VC/TTS systems. During testing, ADTMOS can accurately assess the similarity in the synthesized audio from the same VC/TTS system based on the prior knowledge acquired during training without knowing the system ID.

For two proposed data augmentation methods, ADTMOS-SI-ID achieves better performance in utterance-level LCC, SRCC, and MSE, the values of which are 0.794, 0.729, and 0.393, respectively. This could be attributed to the ID augmentation's ability to effectively expand the sample size without changing the distribution of the training set. However, the system-level MSE for ADTMOS-SI-ID is 0.021, which is higher than ADTMOS-SI's 0.016, indicating that the improvement in system-level prediction performance for ADTMOS-ID is insignificant. One possible reason is that ID data augmentation ensures the distribution consistency of MOS scores between the training and testing sets. However, it alters the distribution of audio samples among VC/TTS systems, affecting the model's accuracy in predicting individual VC/TTS systems. Contrary to ADTMOS-SI-ID, the ADTMOS-SI-PA model, which utilizes PA data augmentation, significantly reduces performance. This is similar to the conclusion in [54], which states that the distribution difference between the training and testing sets greatly impacts prediction accuracy. Therefore, the distribution consistency before and after data augmentation is important.

2) Performance Evaluation on BVCC dataset: Table III shows the experimental results of ADTMOS variants and other baseline models on the BVCC dataset. The ADTMOS-SI-LI model demonstrates superior performance compared to all the other models listed. ADTMOS-SI-LI exhibits a consistently high level of performance on both the BVCC and VCC2018-CSMSC datasets, reaching the utterance-level SRCC of **0.807** and the utterance-level MSE of **0.301**. In particular, ADTMOS-SI-LI outperforms LDNet-MN by 2.21% in utterance-level LCC, 2.35% in utterance-level SRCC, 8.52% in utterance-level MSE, and 3.88% in utterance-level KTAU. Such results demonstrate the outstanding generalization ability of ADTMOS-SI-LI.

In addition, the results between ADTMOS-SI and ADTMOS-SI-ID reveal that incorporating ID data augmentation methods in ADTMOS can certainly improve both the utterance-level and the system-level predicting accuracy, which is consistent with the results in the VCC2018-CSMSC dataset. It indicates that the proposed ID data augmentation method benefits the model's predictive performance across different datasets. Similarly, the PA data augmentation method significantly reduced the predictive accuracy of ADTMOS-SI, again emphasizing the importance of ensuring label distribution consistency between the training and test sets for this task.

An interesting observation is that, unlike the results in the VCC2018-CSMSC dataset where LDNet-ML outperforms LDNet-MN, in the BVCC dataset, the predictive performance of LDNet-ML is worse than LDNet-MN. This may be because the BVCC dataset contains significantly more systems than the VCC2018-CSMSC dataset. Using LD modeling alone cannot capture the audio variances between different systems. The performance impact of this difference becomes more pronounced as the number of systems in the dataset increases.

3) In-depth Analysis of proposed ADT: To further demonstrate the effectiveness of the proposed ADT in ADTMOS for SQA tasks, we assessed the sensitivity and accuracy of different models in predicting distortion by introducing two levels of distortions (mild and severe) to audio generated by different systems. Specifically, we refer to [44] to determine the types of distortions to be applied in the experiment. By adding these distortions to the clean speeches, we can obtain a set of speeches with controlled distortions. The following distortion types are considered:

- **Gain Distortion**: This involves increasing the amplitude of the audio signal by a predefined gain factor, thus amplifying the overall loudness. When the gain exceeds the systems threshold, it results in both time-domain and frequency-domain distortions.
- **White Noise**: Random noise is introduced to simulate background interference often encountered in VC/TTS systems, thereby degrading speech quality.
- **Echo**: This distortion reduces intelligibility by altering the perceived distance and depth of the sound.
- **Clipping Distortion**: Clipping causes the top and bottom of the audio waveform to flatten, leading to both waveform and spectral distortion.

We use 1,780 clear audio samples from the CSMSC dataset and an unseen system (ESPnet-transformerv1) from the BVCC test set. We apply the four distortion types to generate audio with two levels of distortions (mild and severe) by adjusting distortion parameters. We then compare the MOS prediction results of four SQA models, namely MBNet, LDNet, ADTMOS-SI without ADT, and ADTMOS-SI, on audio with different levels of distortions. The results are shown in Fig. 5.

TABLE III
THE UTTERANCE-LEVEL AND THE SYSTEM-LEVEL EXPERIMENTAL RESULTS ON THE BVCC DATASET.

Method	utterance-level							system-level						
	LCC ↑	SRCC ↑	MSE ↓	KTAU ↑	MAE ↓	R2 ↑	MSA_1↑	LCC ↑	SRCC ↑	MSE ↓	KTAU ↑	MAE ↓	R2 ↑	$MSA_{0.5}$↑
MOSNet [21]	0.721	0.712	0.409	0.530	0.512	0.517	88.85%	0.808	0.806	0.243	0.612	0.386	0.626	94.37%
MBNet-MN [24]	0.740	0.746	0.441	0.563	0.518	0.479	87.72%	0.828	0.833	0.226	0.649	0.352	0.653	95.87%
LDNet-MN [25]	0.793	0.789	0.329	0.601	0.456	0.612	91.75%	0.866	0.868	0.196	0.687	0.333	0.698	95.68%
LDNet-ML [25]	0.767	0.757	0.356	0.570	0.479	0.580	90.02%	0.8478	0.841	0.211	0.666	0.353	0.676	95.59%
ADTMOS-SI	**0.812**	0.806	0.304	0.622	0.433	0.641	93.12%	0.911	0.905	0.128	**0.739**	0.275	0.803	97.89%
ADTMOS-SI-LI	0.811	0.807	0.301	**0.624**	**0.427**	0.645	93.53%	0.908	0.901	**0.115**	0.738	**0.256**	**0.823**	**98.03**%
ADTMOS-SI-ID	0.812	**0.808**	**0.296**	0.622	0.433	**0.651**	**94.18**%	**0.911**	**0.907**	0.120	0.738	0.269	0.816	97.56%
ADTMOS-SI-PA	0.784	0.777	0.352	0.594	0.464	0.584	90.43%	0.874	0.863	0.162	0.696	0.297	0.751	96.31%

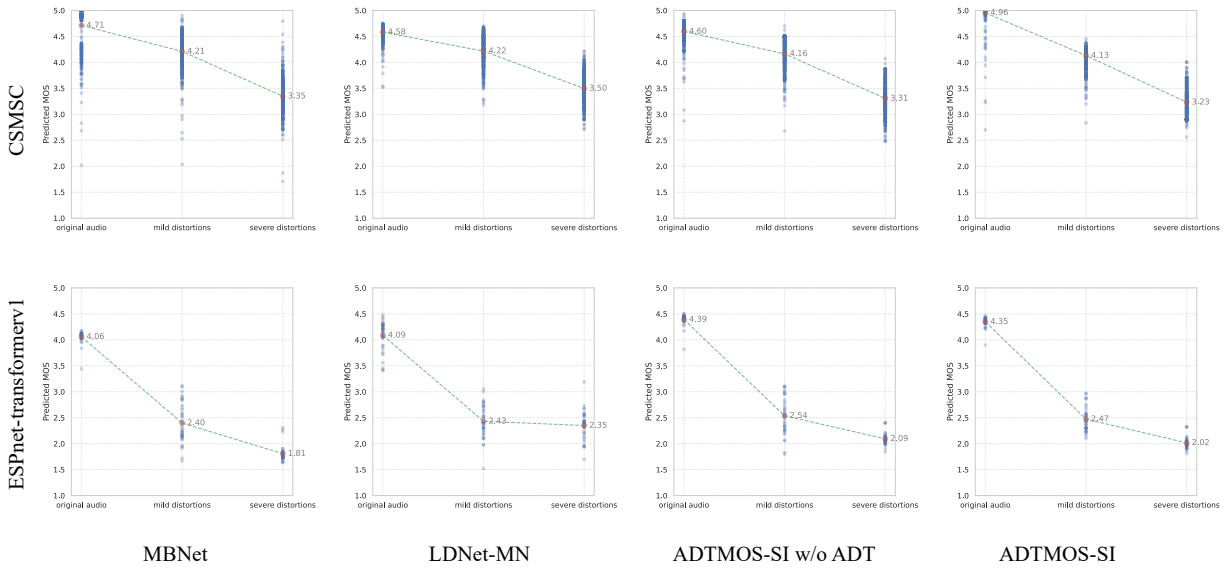


Fig. 5. Scatter plots of MOS prediction results of four different SQA models tested on speeches with different levels of distortion applied. Each column of each plot represents the MOS prediction results for the original audio, audio with mild distortions added, and audio with severe distortions added, respectively. The red dots in the graph indicate the mean predicted MOS scores for each set of audio samples.

Among the 1,780 audio samples with a MOS score of 5 in the CSMSC dataset, the MOS predictions from all models exhibit a decreasing trend. However, the predicted MOS values of ADTMOS-SI are more centralized with the smallest variance, and they are closer to the groundtruth compared to ADTMOS-SI without the ADT for clear audio. For distorted audio, the predicted MOS values of ADTMOS-SI show a downward trend with smaller variance, especially for audio with lower MOS scores This indicates that the ADT enhances the models ability to accurately perceive different types of audio distortions.

Tested using the audio synthesized by the ESPnet-transformerv1 system in the BVCC dataset, all models except LDNet-MN show a decreasing trend in MOS predictions for samples applied two levels of distortion. It means that, when evaluating the MOS value for audio generated by an unknown system, other three models can perceive the severity of distortions residing in it. Additionally, the predicted MOS values of ADTMOS-SI are closest to the groundtruth for original audio, with the most concentrated predictions. By contrast, although MBNet displays a clear decreasing trend, it exhibits high variance in its predictions, which indicates its instability in distortion perception. Its likely due to the influence of content and linguistic characteristics in different audio samples, which MBNet struggles to handle when assessing distortion levels. In the end, LDNet-MN outputs high MOS values for severely distorted audio without an obvious downward trend compared with other three models. The worse performance of LDNet-MN could be due to its over-reliance on listener information and the failure to capture the underlying characteristics of audio distortions.

E. Ablation Study

1) Effect of different modules in ADTMOS: A detailed ablation study was conducted on the VCC2018-CSMSC dataset to verify the effectiveness of different modules in ADTMOS-SI, with the results presented in Table IV.

- **Mix Acoustic Features (MAF)**. After removing the MAF features, ADTMOS-SI drops by 0.51% in system-level SRCC and 44.02% in system-level MSE. It indicates that MAF has a significant improvement effect on the system-level predictive performance. One possible reason is that the concatenation of the audio distortion embeddings χ^{ad} with the system IDs captures the variability of audio distortions between different VC/TTS systems, thus improving the system-level prediction accuracy.
- **MCW**. The MCW module is removed and replaced with a self-attention block [60], and all metrics are signifi-

TABLE IV
ABLATION STUDY OF DIFFERENT MODULES OF ADTMOS ON VCC2018-CSMSC CORPUS.

Method	utterance-level					system-level				
	LCC ↑	SRCC ↑	MSE ↓	KTAU ↑	MSA_1 ↑	LCC ↑	SRCC ↑	MSE ↓	KTAU ↑	$MSA_{0.5}$↑
ADTMOS-SI	**0.789**	0.726	**0.407**	0.560	87.97%	**0.986**	**0.944**	**0.016**	**0.858**	100%
– MAF	**0.789**	**0.728**	**0.407**	**0.562**	87.93%	0.979	0.939	0.023	0.818	100%
– MCW	0.779	0.711	0.416	0.554	87.39%	0.975	0.930	0.030	0.812	100%
– smooth L1 Loss	0.786	0.720	0.416	0.564	87.67%	0.984	0.941	0.020	0.845	100%
– SSL-based embeddings	0.787	0.720	0.409	0.556	**88.25%**	0.979	0.939	0.025	0.823	100%

TABLE V
ABLATION STUDY OF EIGHT ACOUSTIC FEATURES ON VCC2018-CSMSC DATASETS.

Method	utterance-level					system-level				
	LCC ↑	SRCC ↑	MSE ↓	KTAU ↑	MSA_1 ↑	LCC ↑	SRCC ↑	MSE ↓	KTAU ↑	$MSA_{0.5}$↑
ADTMOS-SI	0.789	0.726	0.407	0.560	**87.97%**	**0.986**	**0.944**	**0.016**	**0.858**	100%
– Zero Crossing Rate	0.791	0.726	**0.405**	0.563	87.91%	0.984	0.939	0.018	0.845	100%
– Energy	0.785	0.723	0.411	0.561	87.92%	0.979	0.937	0.021	0.823	100%
– Entropy of Energy	0.791	0.727	0.406	**0.564**	87.96%	0.986	**0.944**	0.017	0.850	100%
– Spectral Centroid	0.788	0.725	0.409	0.560	87.92%	0.979	0.941	0.019	0.829	100%
– Spectral Spread	0.788	0.728	0.407	0.561	87.96%	0.984	0.941	0.018	0.841	100%
– Spectral Entropy	**0.792**	0.726	0.406	0.563	87.94%	0.984	0.943	0.019	0.856	100%
– Spectral Flux	0.787	**0.729**	0.410	0.559	87.96%	0.985	0.943	0.019	0.841	100%
– Spectral Rolloff	0.788	0.726	0.408	0.556	87.94%	0.984	0.942	0.019	0.844	100%

cantly decreased. Compared with the attention module, the MCW module is designed by fusing cross-domain features through 1D convolution and channel weighting to obtain audio distortion-aware information with specific granularity. This module captures the perceptual characteristics of audio distortions better than attention mechanisms.

- **Smooth L1 Loss**. This study uses the smooth L1 loss instead of the MSE loss or the clipped MSE loss. After using smooth L1 loss, the utterance-level MSE of the model improved most significantly, from 0.416 to 0.407. This indicates that the smooth L1 loss can indeed alleviate the impact of outliers in the dataset on the gradient update of the model's loss function. Hence, in future research on speech quality evaluation, the smooth L1 loss can be used to replace the original MSE loss.
- **SSL-based embeddings**. When replacing the SSL-based features with the amplitude spectrum features, a notable decrease is observed in the system-level performance, with the MSE reducing by 55.3%. To explore model performance improvement using features extracted by SSL, we replaced the original amplitude spectrum features with features based on the Wav2Vec2 convolutional layer output proposed by the model in this section's ablation experiment. It shows that using features based on SSL indeed brings more abundant feature information and improves the model performance.

2) Effect of eight mixed acoustic features: An additional ablation study was conducted to evaluate the contribution of mixed acoustic features, including Zero Crossing Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, and Spectral Rolloff. Table V presents the MOS prediction results of the ADTMOS-SI model after removing different acoustic features respectively. Notably, removing either Energy or Spectral Centroid leads to the most pronounced degradation in model performance. This is because Energy directly influences the perceived intensity and loudness of the audio, thereby indirectly affecting its clarity. The spectral centroid, calculated as the weighted average frequency of a signals power spectrum, quantifies the concentration of energy in higher versus lower frequencies (e.g., a higher centroid indicates more high-frequency content). This metric is critical for identifying quality issues such as excessive high-frequency noise or insufficient low-frequency presence in speech signals. Moreover, the performance consistently declines at the system level when any acoustic feature is removed, further underscoring that the integration of mixed acoustic features is essential for the model to accurately capture characteristics of audio distortions and enhance prediction accuracy.

V. CONCLUSION

In this paper, we proposed a novel audio distortion token-guided deep MOS predictor, ADTMOS, which has been highly influential in the SQA task. Our model effectively integrates cross-domain features, including SSL-based audio embeddings, audio metadata, and mixed acoustic features through two sub-networks to simultaneously compute the frame-wise MOS score and the audio distortion score, representing information about different listeners' perceptions of audio distortions. Incorporating audio metadata, including VC/TTS system information and listener information, with SSL-based audio embeddings and employing multichannel weighted fusion mechanisms to extract listeners' perceptual information of audio distortions have further improved the model's performance. Extensive experiments and studies have corroborated the superior performance of ADTMOS compared to several DL-based SQA models.

In addition, to address the problems of lack of data and uneven label distribution in SQA datasets, we propose identically-distributed and proportion-aware data augmentation methods, which expand data samples. Experimental results demonstrate their effectiveness in improving the performance and robustness of SQA models. Moreover, the

impact of different dimensionality unification methods on the predictive performance of SQA models has been further discussed. Detailed comparative experiments show that compared to repetitive padding, zero padding is still a more suitable dimensionality unification method.

REFERENCES

[1] 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.

[2] I. Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, vol. 22, 1996.

[3] I. Recommendation, "562-3, subjective assessment of sound quality," *International Telecommunications Union Radiocommunication Assembly*, 1990.

[4] W. D. Voiers, "Diagnostic acceptability measure for speech communication systems," in *ICASSP 1977*, vol. 2. IEEE, 1977, pp. 204–207.

[5] S. Shirali-Shahreza and G. Penn, "Mos naturalness and the quest for human-like speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 346–352.

[6] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech & Language*, vol. 84, p. 101577, 2024.

[7] V. Grancharov, W. B. Kleijn, J. Benesty, M. M. Sondhi, and Y. A. Huang, "Speech quality assessment," *Springer handbook of speech processing*, pp. 83–100, 2008.

[8] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Fifth international conference on spoken language processing*. ISCA, 1998.

[9] R. Viswanathan, J. Makhoul, and W. Russell, "Towards perceptually consistent measures of spectral distance," in *ICASSP 1976*, vol. 1. IEEE, 1976, pp. 485–488.

[10] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 242–248, 1988.

[11] N. Pourmand, D. Suelzle, V. Parsa, Y. Hu, and P. Loizou, "On the use of bayesian modeling for predicting noise reduction performance," in *ICASSP 2009*. IEEE, 2009, pp. 3873–3876.

[12] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP 2001*, vol. 2. IEEE, 2001, pp. 749–752.

[13] J. G. Beerends, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part itemporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[14] P. Gray, M. Hollier, and R. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 147, no. 6, pp. 493–501, 2000.

[15] I. Union, "Single ended method for objective speech quality assessment in narrow-band telephony applications," *ITU-T Recommendation*, p. 563, 2004.

[16] D.-S. Kim, "Anique: An auditory model for single-ended speech quality estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, 2005.

[17] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[18] R. K. Dubey and A. Kumar, "Non-intrusive objective speech quality assessment using a combination of mfcc, plp and lsf features," in *ICSC 2013*. IEEE, 2013, pp. 297–302.

[19] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, "A hierarchical predictor of synthetic speech naturalness using neural networks," in *Proceedings of Interspeech 2016*. ISCA, 2016, pp. 342–346.

[20] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," in *Proceedings of Interspeech 2018*. ISCA, 2018, pp. 1873–1877.

[21] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning-based objective assessment for voice conversion," in *Proceedings of Interspeech 2019*. ISCA, 2019, pp. 1541–1545.

[22] Y. Choi, Y. Jung, and H. Kim, "Deep mos predictor for synthetic speech using cluster-based modeling," in *Proceedings of Interspeech 2020*. ISCA, 2020, pp. 1743–1747.

[23] A. Vioni, G. Maniati, N. Ellinas, J. S. Sung, I. Hwang, A. Chalamandaris, and P. Tsiakoulis, "Investigating Content-Aware Neural Text-To-Speech MOS Prediction Using Prosodic and Linguistic Features," pp. 1–5, 2023.

[24] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNET: MOS Prediction for Synthesized Speech with Mean-Bias Network," in *ICASSP 2021*. IEEE, 2021, pp. 391–395.

[25] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech," in *ICASSP 2022*. IEEE, 2022, pp. 896–900.

[26] M. Chinen, J. Skoglund, C. K. Reddy, A. Ragano, and A. Hines, "Using rater and system metadata to explain variance in the voicemos challenge 2022 dataset," in *Proceedings of Interspeech 2022*. ISCA, 2022, pp. 4531–4535.

[27] W.-C. Tseng, C.-y. Huang, W.-T. Kao, Y. Y. Lin, and H.-y. Lee, "Utilizing Self-supervised Representations for MOS Prediction," 2021.

[28] X. Tian, K. Fu, S. Gao, Y. Gu, K. Wang, W. Li, and Z. Ma, *A Transfer and Multi-Task Learning Based Approach for MOS Prediction*. ISCA, 2022.

[29] W.-C. Tseng, W.-T. Kao, and H.-y. Lee, "DDOS: A MOS Prediction Framework utilizing Domain Adaptive Pre-training and Distribution of Opinion Scores," pp. 4541–4545, 2022.

[30] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization Ability of MOS Prediction Networks," in *ICASSP 2022*. IEEE, 2022, pp. 8442–8446.

[31] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," in *Proceedings of Interspeech 2022*. ISCA, 2022, pp. 4521–4525.

[32] A. B. Aicha and S. B. Jebara, "Perceptual speech quality measures separating speech distortion and additive noise degradations," *Speech Communication*, vol. 54, no. 4, pp. 517–528, 2012.

[33] Z. Zhang, D. S. Williamson, and Y. Shen, "Investigation of phase distortion on perceived speech quality for hearing-impaired listeners," in *Proceedings of Interspeech 2020*. ISCA, 2020, pp. 2512–2516.

[34] K. H. Arehart, J. M. Kates, M. C. Anderson, and L. O. Harvey, "Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1150–1164, 2007.

[35] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," in *ICASSP'98*, vol. 1. IEEE, 1998, pp. 541–544.

[36] G. Chen and V. Parsa, "Loudness pattern-based speech quality evaluation using bayesian modeling and markov chain monte carlo methods," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. EL77–EL83, 2007.

[37] B. C. Moore, *An introduction to the psychology of hearing*. Emerald, 2012.

[38] L. Ding and R. A. Goubran, "Speech quality prediction in VoIP using the extended E-model," in *GLOBECOM '03. IEEE Global Telecommunications Conference (IEEE Cat. No.03CH37489)*, vol. 7. IEEE, 2003, pp. 3974–3978.

[39] M. Yu, C. Zhang, Y. Xu, S. Zhang, and D. Yu, "Metricnet: Towards improved modeling for non-intrusive speech quality assessment," *arXiv preprint arXiv:2104.01227*, 2021.

[40] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2022.

[41] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[42] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?" in *ICASSP 2021*. IEEE, 2021, pp. 6533–6537.

[43] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Proceedings of Interspeech 2022*. ISCA, 2021, pp. 2278–2282.

[44] A. Wilson and B. Fazenda, "Categorisation of distortion profiles in relation to audio quality," in *17th International Conference on Digital Audio Effects*, 2014.

[45] P. Vieting, R. Schlüter, and H. Ney, "Comparative analysis of the wav2vec 2.0 feature extractor," in *ITG 2023*. IKS, 2023.

[46] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, "Perceptual quality dimensions of text-to-speech systems," in *Proceedings of Interspeech 2011*. ISCA, 2011.

[47] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[48] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 14 136–14 147.

[49] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 5297–5307.

[50] Y. Tang, X. Zhang, L. Ma, J. Wang, S. Chen, and Y.-G. Jiang, "Non-local netvlad encoding for video classification," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. Springer, 2018, pp. 0–0.

[51] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, pp. 1–52, 2020.

[52] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[53] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *The Speaker and Language Recognition Workshop (Odyssey 2018)*. ISCA, 2018, pp. 195–202.

[54] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The voicemos challenge 2022," in *Proceedings of Interspeech 2022*. ISCA, 2022, pp. 4536–4540.

[55] DataBaker Technology Co., Ltd., "Chinese standard mandarin speech corpus," https://www.data-baker.com/open_source.html.

[56] G. R. Terrell and D. W. Scott, "Variable kernel density estimation," *The Annals of Statistics*, vol. 20, no. 3, pp. 1236–1265, 1992.

[57] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *ICASSP 2022*. IEEE, 2022, pp. 6182–6186.

[58] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP 2015*. IEEE, 2015, pp. 5206–5210.

[59] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of ICLR 2019*. ICLR, 2019.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.