aerospace

MDPI

Article

ARCnet: A Multi-Feature-Based Auto Radio Check Model

Weijun Pan, Yidi Wang *, Yumei Zhang and Boyuan Han

College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China; wjpan@cafuc.edu.cn (W.P.); zhanyme@cafuc.edu.cn (Y.Z.); hanregulu@gmail.com (B.H.)
* Correspondence: wyd1192254593@163.com

Abstract: Radio checks serve as the foundation for ground-to-air communication. To integrate machine learning for automated and reliable radio checks, this study introduces an Auto Radio Check network (ARCnet), a novel algorithm for non-intrusive speech quality assessment in civil aviation, addressing the crucial need for dependable ground-to-air communication. By employing a multi-scale feature fusion approach, including the consideration of audio's frequency domain, comprehensibility, and temporal information within the radio check scoring network, ARCnet integrates manually designed features with self-supervised features and utilizes a transformer network to enhance speech segment analysis. Utilizing the NISQA open-source dataset and the proprietary RadioCheckSpeech dataset, ARCnet demonstrates superior performance in predicting speech quality, showing a 12% improvement in both the Pearson correlation coefficient and root mean square error (RMSE) compared to existing models. This research not only highlights the significance of applying multi-scale attributes and deep neural network parameters in speech quality assessment but also emphasizes the crucial role of the temporal network in capturing the nuances of voice data. Through a comprehensive comparison of the ARCnet approach to traditional methods, this study underscores its innovative contribution to enhancing communication efficiency and safety in civil aviation.

Keywords: radio check; speech quality assessment; multi-feature module

Citation: Pan, W.; Wang, Y.; Zhang, Y.; Han, B. ARCnet: A Multi-Feature -Based Auto Radio Check Model. *Aerospace* **2024**, *11*, 391. https://doi.org/10.3390/aerospace11050391

Academic Editor: Guanjun Xu

Received: 16 April 2024
Revised: 11 May 2024
Accepted: 13 May 2024
Published: 14 May 2024

1. Introduction

To ensure flight safety and efficiency in the field of civil aviation, effective communication between aerodromes and airspace must be established. The very-high-frequency communication system (VHF COMM) is a vital component in the realm of mobile wireless communication and plays a pivotal role in the communication of civil aviation. Operating primarily through voice transmission, the VHF COMM system employs very-high-frequency radio signals to convey information between communicating parties. Air traffic controllers monitor aircraft in the airspace using radar surveillance within the air traffic control system. Positioned at air traffic control stations, controllers provide airborne traffic services by issuing voice commands through the VHF COMM system, enabled by interphone systems, to aircraft within their assigned airspace sectors.

The very-high-frequency communication system has an operational frequency range of 30 to 300 MHz, placing it within the very-high-frequency spectrum. VHF radios serve as a primary communication tool within the A1 maritime area of the Global Maritime Distress and Safety System (GMDSS), serving as a pivotal means for on-site communication and the sole method for communication between air traffic controllers and pilots. Presently, VHF radios used in civil aviation operate between 118.000 and 151.975 MHz (with an actual maximum frequency of 136 MHz), with a frequency spacing of 25 kHz per channel. This frequency range and channel spacing are stipulated by the International Civil Aviation Organization. Notably, the frequencies between 121.600 and 121.925 MHz are designated primarily for ground control.

The very-high-frequency (VHF) system operates using amplitude modulation (AM) and has a minimum transmitter output power of 20 W. The key characteristics of VHF

transmission are as follows: Due to their high frequency, surface waves experience rapid attenuation, resulting in short propagation distances along the line of sight. As a result, VHF communication relies primarily on space wave propagation, which is highly susceptible to tropospheric effects. Additionally, terrain and geographical features have a significant impact. Factors such as weather, encoders, and thermal noise within the radiotelephone communications channel often lead to challenges such as high noise levels and voice signal distortion in communications.

The VHF radiotelephone communication system is the primary communication network for commercial aircraft within the civil aviation system. It enables bidirectional voice communication between the flight crew and the air traffic controller at various stages of flight. The takeoff and landing phases are the busiest periods for aircraft pilots, and also the most susceptible to accidents. To ensure accurate information exchange between air traffic controllers and flight crews during these crucial phases, VHF communication must be highly reliable. Therefore, real-time evaluation and monitoring of the communication quality between ground and airborne parties becomes a focal point of radiotelephone communications in civil aviation.

In radiotelephone communications, the "Radio Check" command is typically used to determine the operational status and signal quality of voice transceivers and to establish contact with ground control stations. The "Radio Check" command pertains to radio verification and evaluates the quality and consistency of radio signals. In aerial communications, pilots communicate with air traffic controllers to inquire about the clarity of the signals they are receiving. The controller responds and provides advice to ensure smooth communication. According to ICAO's official recommendations, the sequence for a radio verification is as follows: the other party's call sign + one's own call sign + "Radio Check" + "how do you read". In response, the signal quality is categorized into five levels and reported as follows: Unreadable, Sometimes readable, Readable but with difficulty, Readable, and Perfectly readable [1–3].

As shown in Figure 1, when an aircraft initially enters an airspace, it makes contact with an air traffic controller and executes the "Radio Check" command to conduct a radio signal evaluation. The pilot's voice command undergoes modulation through a very high frequency (VHF) before transmission from the aircraft. Upon reception by the aerodrome's VHF equipment, the signal is routed to the interphone systems. Subsequently, the air traffic controller evaluates the speech quality based on the received signal, and the evaluation outcome aids both parties in fine-tuning their equipment.

In the actual communication environment of airports, the following reasons may lead to the deterioration of communication, requiring the parties involved to perform a radio check:

1. Poor weather conditions, which often introduce some noise.
2. Incorrect use of the transmitter by the communicating parties, such as using mismatched frequency bands (which is generally the main reason for conducting a radio check), and placing the transmitter too close to the speaker's mouth (this can cause popping sounds, which is also a secondary reason for conducting a radio check).
3. Aging communication equipment. The current very-high-frequency communication system, compared to the latest WLAN or cellular communications, still modulates, transmits, and demodulates the original signal. The electronic components filter the signal which can lead to a loss of information and the introduction of noise.

In the actual operation of airports, such interactions are plentiful, and effective radio communication can enhance the efficiency of airport operations [4]. Under poor communication conditions, both ground and air parties expend significant time and effort on radio adjustments, which, on average, reduces the efficiency of airport operations and poses safety risks [5]. We have analyzed dialogue data from several major airports in China. During the actual operation of airports, the time taken for one command exchange is about 10 s. Including the time taken by both parties to adjust the transmitter, one manual radio adjustment takes about 10 to 20 s, while the actual duration of a normal conversa-

tion does not exceed one minute. Moreover, currently, in the actual operation of airports, both parties must manually perform radio checks. Therefore, employing machines for radio adjustments can significantly increase the information throughput of ground-to-air communications, thereby improving the efficiency of controllers.

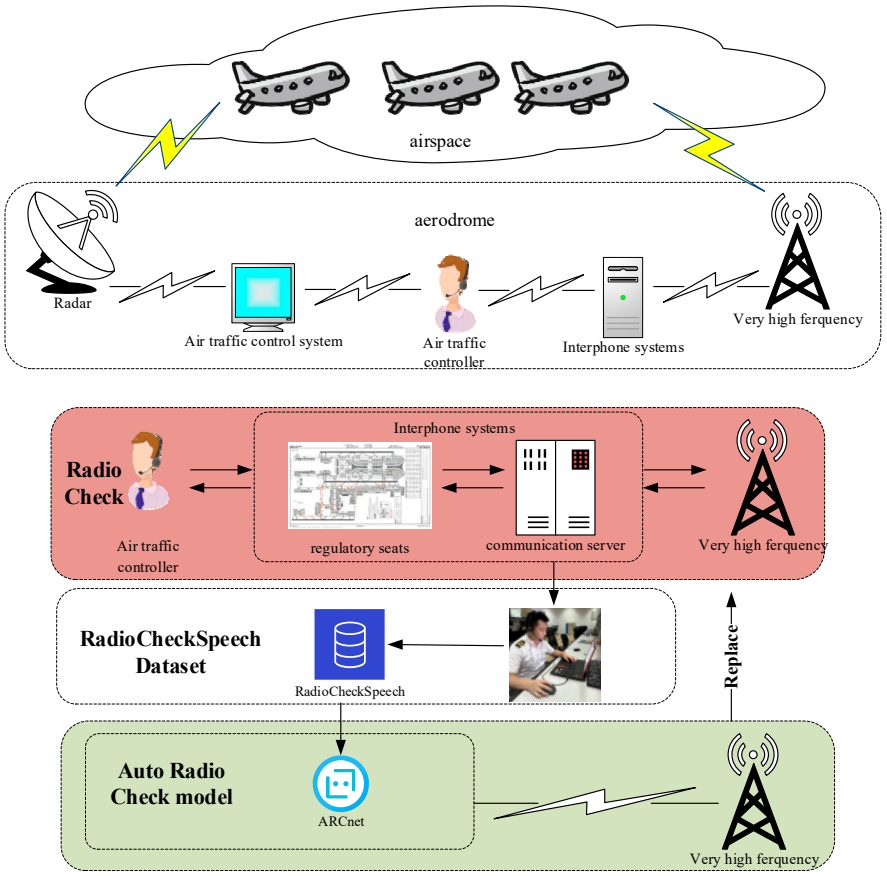


Figure 1. This is a sample of a figure caption.

This study aims to replace the manual "Radio Check" scoring process with an automated machine-based evaluation. By harnessing voice data received by the aerodrome's VHF equipment, the Auto Radio Check model emulates the controller's perception of speech quality. The real-time assessment of voice data is conducted using the results of the "Radio Check" scoring. This method not only streamlines the cumbersome radio verification procedure performed each time communication is established between ground and airborne parties, thereby improving communication efficiency, but it also provides a more immediate and intuitive depiction of the channel quality.

Hence, the objective of this study is to compile a RadioCheckSpeech Dataset consisting of radiotelephone communication voice data obtained from interphone systems and evaluated by air traffic controllers. This dataset will be used to train and validate models.

The evaluation of voice in radiotelephone communication heavily relies on the most immediate subjective perception of air traffic controllers, similar to the mean opinion score (MOS) assessment method. MOS is a subjective measurement method used to evaluate the quality of speech. It involves obtaining subjective ratings for voice samples from a group of individuals, including experts and non-experts. MOS scores are widely used in the telecommunications industry to assess call quality [6–8]. The MOS and the "Radio Check" command in radiotelephone communication both share similarities as subjective, non-intrusive methods for voice assessment. A comparison of their evaluation methods is presented in Table 1 below:

Table 1. Radio check compared to mean opinion score.

Score	Quality	Radio Check	Mean Opinion Score	Score
1	Bad	Unreadable	Very Annoying	1
2	Poor	Readable now and then	Annoying	2
3	Fair	Readable but with difficulty	Slightly annoying	3
4	Good	Readable	Perceptible but not annoying	4
5	Excellent	Perfectly readable	Imperceptible	5

Currently, researchers are conducting extensive research on deep-learning-based speech quality prediction models, with the goal of replacing manual perceptual assessments with more accurate and cost-efficient methods. These efforts have accumulated numerous models and achievements. Intrusive speech quality assessment algorithms and non-intrusive speech quality assessment algorithms are the two most common classifications for such speech quality assessment algorithms. The distinction lies in the fact that intrusive algorithms require clean reference signals, whereas non-intrusive algorithms do not.

Non-intrusive speech quality assessment algorithms can evaluate speech quality without requiring reference signals. Due to the requirement for pristine, noise-free reference speech in intrusive assessments, non-intrusive algorithms are increasingly utilized in real-world settings. For instance, Lo et al. proposed MOSNet, a MOS predictor based on convolutional neural networks and bidirectional long short-term memory networks, which is used to predict MOS scores of converted speech in speech transformation systems [9]. Fu et al. introduced Quality-Net, a speech quality assessment model based on bidirectional long short-term memory networks, which is used to predict MOS scores of enhanced speech generated by speech enhancement systems [10]. Yoshimura et al. presented a synthetic speech naturalness predictor based on fully connected neural networks and convolutional neural networks, which is used to predict MOS scores of synthetic speech generated by speech synthesis systems [11]. Naderi et al. presented AutoMOS, a naturalness assessment model based on autoencoders and convolutional neural networks, for predicting MOS scores of synthetic speech generated by speech synthesis systems [12]. Mittal et al. proposed NISQA, a non-intrusive speech quality assessment model based on convolutional neural networks and long short-term memory networks, for predicting MOS scores of speech with various channel noise introductions [13]. ITU-T Rec. P.563 is a single-ended method for objective speech quality assessment in narrowband telephone applications, approved in 2004 [14]. ANIQUE+ is a novel national standard used in the United States for non-intrusive estimation of narrowband speech quality [15]. A. A. Catellier and S. D. Voran introduced WAWEnets, a reference-free convolutional waveform method for estimating narrowband and wideband speech quality [16].

Some examples of intrusive speech quality assessment algorithms for comparison are given as follows: ITU-T Rec. P.863 is a perceptual objective listening quality assessment method, approved in 2018 [7]. The Integral and Diagnostic Intrusive Prediction of Speech Quality is an algorithm proposed by N. Coté that employs dual-end comparison for speech quality evaluation [17]. M. Chinen et al. introduced the open-source, production-ready speech and audio metric ViSQOL v3 [18].

As mentioned above, statistical speech assessment algorithms like P503, although more interpretable than deep-learning-based speech assessment algorithms, perform poorly in radio check evaluation tasks. In experiments, it was found that statistical speech assessment algorithms have a significant gap in perception compared to actual evaluators of speech quality. Therefore, the ARCnet network still opted for a deep-learning-based network architecture. Moreover, in practical research, it was discovered that controllers performing radio check scoring not only focus on the impact of noise in the speech but also on its comprehensibility. Speech quality assessment algorithms such as MOSNet, AutoMOS, and NISQA primarily focus on noise, with the models using mel-frequency cepstrum coefficient features as the vector for extracting speech quality. The NOMAM model and the speech

evaluation algorithm proposed by Fu et al. utilize self-supervised learning features for assessing speech quality, but the self-supervised vector training mentioned still focuses on extracting speech noise, using noise characterization to predict speech quality [19,20]. Therefore, in the design of ARCnet, not only were mel-frequency cepstrum coefficient features strongly correlated with noise used, but self-supervised vector representations for comprehensibility features relevant to downstream tasks like speech recognition were also considered. In the multi-feature module network design, simply concatenating two different features is insufficient. As the research by Liu et al. suggests, skillfully fusing tensors can enhance the network's perception of speech quality [21]. Therefore, in the design of the ARCnet network, we perform differentiated information extraction for each type of feature, then concatenate them, and in subsequent ablation experiments, this network design approach has been shown to improve network performance.

In summary, Currently, traditional voice evaluation algorithms predominantly process speech using spectral features, focusing solely on the impact of audio noise on speech quality. However, in evaluating radio check scores, the intelligibility of the speech within the audio, that is, whether the audio can be accurately transcribed into text commands, must also be considered. This aspect of speech intelligibility serves as one of the dimensions for controllers when scoring a radio check. Therefore, one of the challenges in designing a radio check scoring network lies in encapsulating the audio's spectral, intelligibility, and temporal information. The current speech evaluation algorithms rely heavily on mean opinion score (MOS) ratings as the primary evaluative metric. These algorithms are often applied in voice over Internet protocol (VOIP) networks, where voice data is transmitted in the form of data packets. In this context, factors affecting speech quality include network-related issues such as packet loss and latency. However, within the realm of civilian aviation communication, evaluation is typically conducted through radio check scoring. In this scenario, very-high-frequency communication systems directly modulate and demodulate voice signals using radio waves. Influential factors on voice quality encompass channel noise and interference generated by components. Currently, AI-powered networks have reached a significant level of maturity in predicting speech quality using the MOS scoring method. Although the MOS scoring method and the radio check scoring method produce similar results, there has been no prior research on the use of AI-based algorithms for evaluating radiotelephone communication speech quality in civil aviation communication systems. Given this context, the major contributions of this paper are as follows:

1. We present for the first time a non-intrusive speech quality assessment algorithm based on the radio check standard for radiotelephone communication that combines human-engineered and self-supervised features. On the NISQA dataset and our proprietary RadioCheckSpeech dataset, comparative evaluations against state-of-the-art speech assessment algorithms were performed. The proposed algorithm demonstrated relative performance enhancements, achieving a 6% increase in Pearson correlation coefficient and a 12% reduction in root mean square error (RMSE) on the NISQA dataset, as well as a 12% increase in Pearson correlation coefficient and a 12% reduction in RMSE on the RadioCheckSpeech dataset.
2. In this study, a dataset titled "RadioCheckSpeech" containing authentic voice commands recorded from internal communication systems at various Chinese airports, control units, and control systems was established. The research employed a method where air traffic controllers assessed these voice commands using the "Radio Check" procedure. Professional controller ratings were collected and manually verified in order to compare mean opinion score (MOS) ratings to radio check scores. Additionally, this dataset was utilized for the network to learn how controllers perform radio check evaluations on speech. The dataset consists of 3200 voice commands spoken in a combination of Chinese and English.

In the second section of the article, the experiments primarily focused on comparing the similarities and differences between MOS ratings and radio check scores. The third section addressed the differences between these two rating methods and introduced the

RadioCheckSpeech dataset. The fourth section describes the design of a network named ARCnet, which combines human-engineered and self-supervised features. The fifth section encompassed comparative experiments between ARCnet and other open-source speech quality prediction models on both the NISQA dataset and the RadioCheckSpeech dataset. Additionally, the section included an ablation analysis of ARCnet's features. Finally, in the sixth section, the article provided a summary of its content.

2. Preliminary Analysis

Based on the introduction in the first chapter, it is evident that MOS scores and radio check scores share a remarkable similarity in terms of their scoring methodologies. Given this similarity, the question arises as to whether an algorithm designed to predict MOS scores could also predict radio check scores. Therefore, it is essential to design experiments that confirm the similarity between these two scoring methods. In the experiments, we opted to use a publicly available, pre-labeled dataset from NISQA with MOS scores [11]. For this dataset, we engaged air traffic controllers to provide radio check scores as well. We divided the dataset into five intervals based on MOS scores between 0 and 5. Within each interval, we randomly selected 40 samples, amounting to a total of 200 samples. According to radio check scoring rules, air traffic controllers then assigned scores to these samples. The resulting scores were fitted with a second-order polynomial, the results of which are depicted in Figure 2 below:

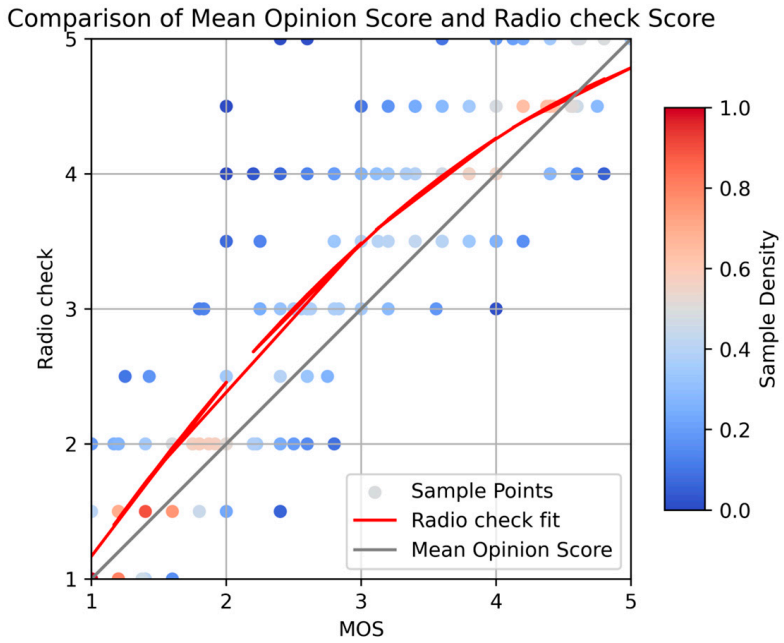


Figure 2. Comparison of mean opinion score and radio check score.

Observing the results, it is evident that MOS scores and radio check scores share a significant trend similarity. In situations where the speech quality is either poor or excellent, the sample density is higher, and this is represented by the red region of the graph, indicating that the two scoring methods agree closely. However, in certain instances where the speech quality is of moderate levels, there is a noticeable discrepancy between the two methods, represented by the blue region. Moreover, radio check scores are typically higher than average. This divergence could potentially be attributed to the differences in how very-high-frequency (VHF) communication networks and voice over Internet protocol networks handle voice data, as well as the inherent dissimilarities between the radio check scoring method and other scoring approaches. To gain a deeper understanding, we conducted a thorough analysis of a voice sample extracted from authentic radiotelephone communication at Suilin Anju Airport.

As shown in Figure 3, the red-bordered section represents the pilot's voice, whereas the green-bordered section represents the air traffic controller's voice. Analyzing the waveform and spectrogram, it is evident that in comparison to the clear voice from the air traffic controller's end, the pilot's voice contains a low-frequency noise at approximately 1000 Hz. This continuous noise on the pilot's side causes the voice envelope to become less distinct and challenging to discern. In the spectrogram, the voice of the pilot has been band-pass filtered between 30 Hz and 3000 Hz, resulting in the loss of both high- and low-frequency features. Additionally, the audio waveform from the pilot's end is relatively weaker, which severely degrades the voice communication quality.

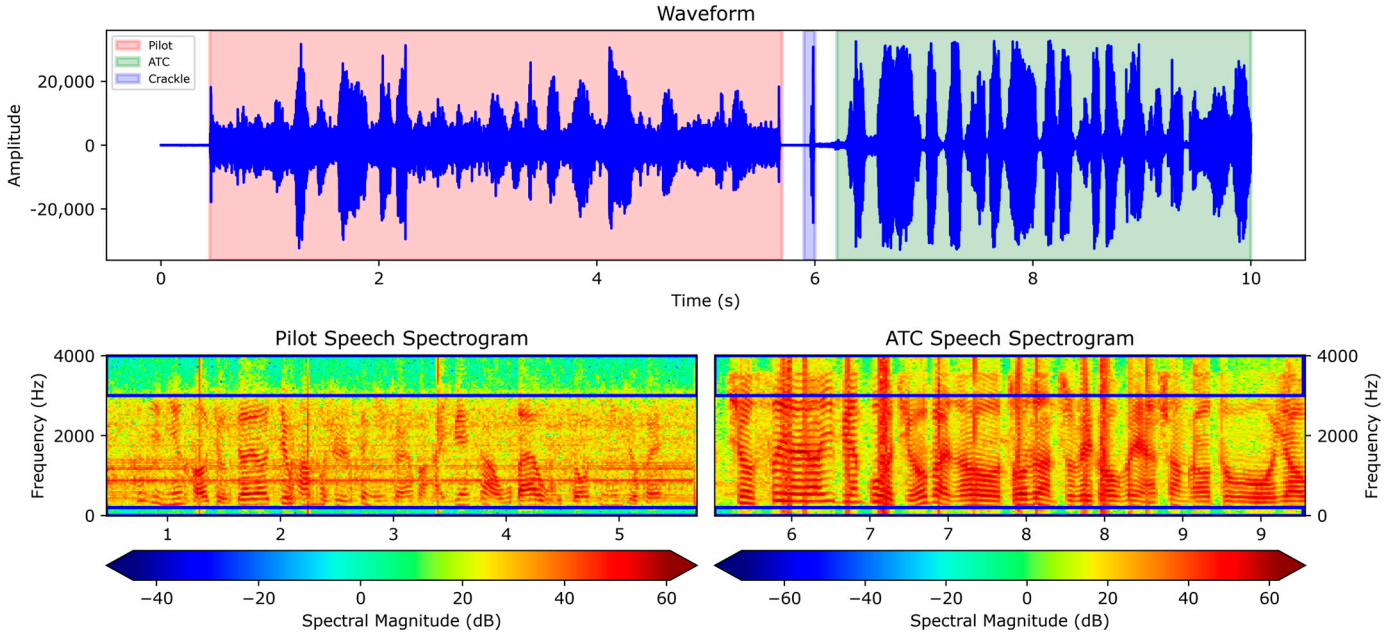


Figure 3. Acoustic and spectral analysis of actual radiotelephone communication at Suilin Anju Airport.

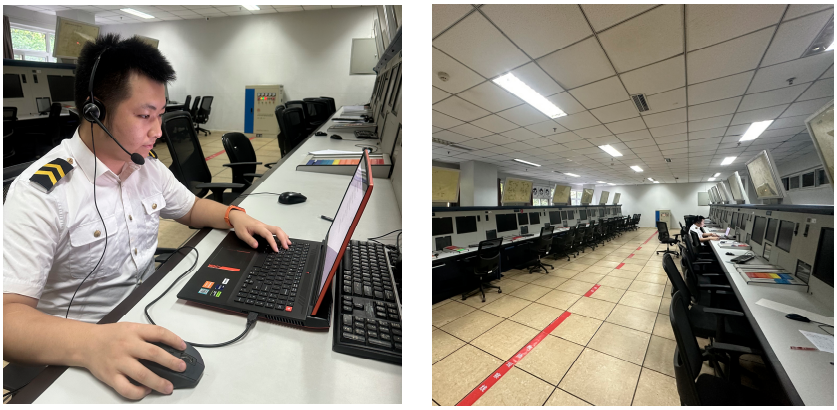
The fundamental reason behind this phenomenon lies in the distinct processing approaches of the very-high-frequency communication system (VHF COMM) used in radiotelephone communication as compared to the processing methods of voice over Internet protocol systems. This divergence diminishes the quality of speech. Therefore, during radio check training, instructors emphasize that the most important aspect of evaluating speech quality is the clarity of command comprehension. If the command is clear and understandable, it is rated a 5. If the command is unclear, it is rated a 4 or 3. However, if the command is difficult to comprehend, it may receive a rating of 2 or 1. Nevertheless, the lack of well-defined standards for scoring between 4 and 3, as well as between 2 and 1, may result in inconsistent evaluations, with air traffic controllers possibly awarding a score of 3 or 4 for the same voice command.

This implies that radio check evaluations not only focus on the influence of noise but also consider the intelligibility of the voice itself. Therefore, for the selection of network features, a multi-feature fusion strategy was utilized, which included the extraction of both human-engineered voice features and self-supervised features. Human-engineered voice features are used to capture noise-related perceptual characteristics affecting speech quality, whereas self-supervised voice features are used to capture features affecting voice intelligibility. This indicates that MOS machine scoring model insights can be applied to radio check scoring research. However, despite the similarities between radio check and MOS scoring methods and outcomes, slight differences still exist. Therefore, in the next section, a dedicated dataset is established for the development of a self-supervised model for radio check scoring.

3. RadioCheckSpeech Dataset

In this section, the RadioCheckSpeech training and evaluation dataset was created.

The RadioCheckSpeech dataset was developed using voice data evaluated by licensed air traffic controllers with at least two years of experience at airports. The voice data originates from several major airports in China, copied from the intercom systems at different control stages (approach, area control, tower), such as (Wuhan Tianhe Airport, Nanba Airport, etc.). The evaluation method resembles the radio check technique utilized during ground-to-air communication. To replicate the scenario in which actual air traffic controllers evaluate speech quality from the aerodrome control tower, we arranged for controllers to evaluate each command in a quiet control simulation environment while wearing specialized headphones. As shown in Figure 4a,b.



(**a**) (**b**)

Figure 4. Air traffic controllers rate the voice using radio check evaluation. (**a**) Shows the controller performing a Radio Check in a near-field environment. (**b**) Shows the controller performing a Radio Check in a far-field environment.

This data set consists of two main components totaling 3200 voice samples with radio check labels. The first component comprises 200 voice samples from the NISQA dataset that were initially labeled with MOS scores. These samples were re-evaluated by air traffic controllers using the radio check scoring method. The second component consists of 3000 authentic radiotelephone communications voice samples recorded in various Chinese regions and airports. This collection contains Chinese and English commands spoken at various speeds, all of which were obtained from the internal communication system. Each voice sample represents a single radiotelephone communication command and has been evaluated by air traffic controllers using the radio check scoring system.

4. Proposed Framework

In this study, we present a fusion model that combines self-supervised learning, human-engineered mel spectrogram features, and a transformer-based feature extraction backbone. This has led to the creation of a multi-feature fusion model. Figure 5 below illustrates the proposed network architecture.

Specifically, the input speech is segmented initially. Each segment of the speech is then passed through the multi-feature module for feature extraction. Self-supervised feature extraction and mel spectrogram feature extraction are involved. The extracted features are concatenated, and then processed using a transformer-based feature extraction model. Finally, the output is the radio check score.

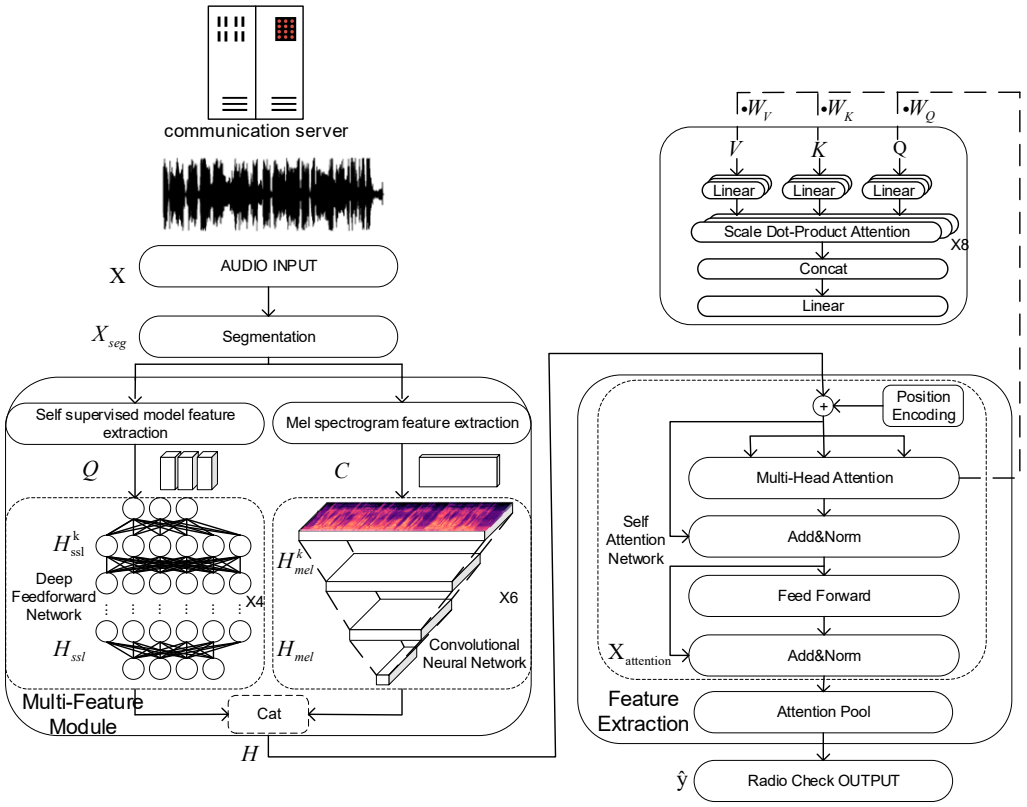
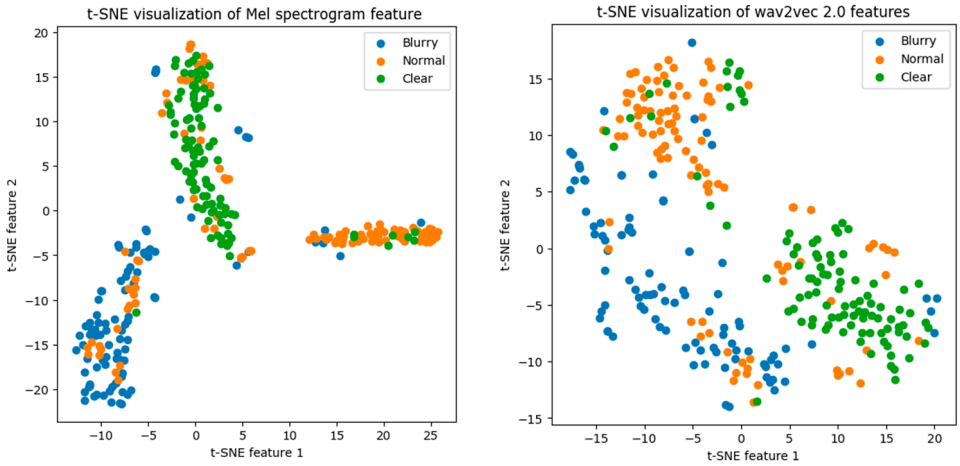


Figure 5. Backbone network structure diagram.

4.1. Feature Interpretability Discussion

In the design of the ARCnet network, the most crucial step is selecting an appropriate method for encoding audio. The chosen audio encoding features should be as relevant as possible to the target task and possess feature interpretability. As described in the Preliminary Analysis section of the document, to enable the ARCnet network to perceive content related to noise and comprehensibility in speech, the selected encoding features are mel spectrogram features and wav2vec2.0 features. To validate the relationship between these features and speech quality, an experiment was conducted. A batch of audio was played for multiple participants to listen to and rate. The evaluations were based on two dimensions: the clarity of the text in the audio and the level of noise. The rating scale was as follows: 1: Burry, 2: Normal, 3: Clear. After evaluation, 100 audio samples were selected from each rating category according to the same scoring dimension to form a noise perception dataset and a comprehensibility perception dataset. In the noise perception dataset, all audios were encoded using mel spectrogram features, while in the comprehensibility perception dataset, audios were encoded using wav2vec2.0 features. After encoding, the t-SNE (t-distributed stochastic neighbor embedding) method was used to analyze both datasets separately, with the analysis results shown in Figure 6a,b.

t-SNE is a statistical method used to reduce high-dimensional data to two or three dimensions for easier visualization. It operates by preserving the relative proximity of similar points in the original high-dimensional space, thus reflecting the structural features of high-dimensional data in the reduced lower-dimensional space. t-SNE excels at maintaining the local structure of data points, meaning that points close to each other in a high-dimensional space remain close in a reduced space. Additionally, t-SNE can reveal clustering structures within the data, even without explicitly using clustering algorithms. Since the t-SNE method lacks the learnability of neural networks, it leads to an interpretative dimension reduction in data. Therefore, analyzing encoded speech data with t-SNE provides an intuitive understanding of whether the encoding method can represent the data's features, verifying the relevance of encoding features to the target task.



(**a**) (**b**)

Figure 6. t-SNE visualization of mel spectrogram feature and wav2vec2.0 feature. (**a**) Shows the t-SNE visualization of Mel spectrogram features. (**b**) Shows the t-SNE visualization of wav2vec 2.0 features.

As shown in Figure 6a, after reducing the data from the noise perception dataset encoded with mel spectrogram features to two dimensions using t-SNE, the data naturally clusters into three segments. This indicates that data encoded with mel spectrogram features are significant, and using mel spectrogram features can well represent the characteristic of audio noise. However, the presence of some different types of points in various segments suggests that more complex neural networks should be used for further feature extraction after encoding with mel spectrogram features, as using only mel spectrogram features is not precise enough.

As shown in Figure 6b, after encoding the comprehensibility perception dataset data with wav2vec2.0 features and reducing it to two dimensions using t-SNE, the data clusters according to evaluation metrics, although it does not cluster into three distinct segments. The enrichment of data under the same labels indicates that wav2vec2.0 features have a strong correlation with speech comprehensibility (ease of understanding spoken words). However, the unclear boundaries between clusters suggest that linear networks should be used to further enhance the features of wav2vec2.0, to achieve better perception.

In summary, mel spectrogram features effectively represent noise but require further refinement through more complex neural networks. wav2vec2.0 features strongly correlate with speech comprehensibility (clarity of spoken words), but still need linear networks to enhance the features. Both noise and speech comprehensibility directly impact the evaluation results of radio checks. Therefore, in the design of ARCnet, mel spectrogram features and wav2vec2.0 features will be used to encode speech.

4.2. Method

In the design of the ARCnet network, the foremost consideration is which voice representation method to use for characterizing speech. The experiment description in the "Preliminary Analysis" section states that during the actual radio check process, evaluators focus more on the impact of noise on speech quality and whether the speech content can be fully transcribed. As shown in Figure 3 of the "Preliminary Analysis", the presence of noise in radio communication is mainly indicated by the long solid lines in the spectrogram, as well as some shapes with graphic features. Therefore, using image recognition methods with convolutional neural networks can easily extract the masking features of noise on the original audio, allowing the model to understand the impact of noise on audio through neural network learning. Secondly, the ability to fully transcribe the textual content boils down to a speech recognition task. In speech recognition tasks, researchers have developed a novel feature, self-supervised learning features, to project speech into a

higher-dimensional textual domain, enabling the neural network to recognize speech as text. When a segment of speech is clear, a hyperplane can be used to separate the speech projected in the textual domain. Conversely, if the speech is noisy and indistinguishable, the corresponding hyperplane is twisted and complex. To allow the network to perceive whether the textual content in speech can be fully transcribed, a linear network processes the projected vectors. The linear network can be understood as a trainable hyperplane, and ultimately, the recognizability of text in speech can be determined by observing the output of the hyperplane.

The network operates as follows: Initially, the audio is segmented, and a Self-supervised feature model is employed to obtain features indicative of the speech's textual domain comprehensibility confidence. Commonly utilized in upstream speech recognition tasks, the Self-supervised feature model maps audio into vectors that neural networks can understand, thus facilitating the translation from the audio domain to the textual domain. Additionally, this model can output a comprehensibility confidence interval for the current audio, indicating whether the audio segment can be accurately transcribed. Through the mapping by the Self-supervised feature network, features related to the speech's textual domain comprehensibility confidence are obtained. Although this process yields a vector feature of speech's textual comprehensibility, these features are initially coarse. A deep feed-forward network is then used to refine these features, producing a more defined confidence feature regarding the segment's textual comprehensibility. Similarly, mel spectrogram features are used to obtain the audio segment's frequency domain vectors. To capture features of noise in the frequency domain that affect audio quality, the network employs a convolutional neural network for noise feature extraction. Pre-trained convolutional layers are sensitive to discontinuities (commonly caused by communication interference), long solid lines (typically resulting from noise), and unnatural envelopes in the spectrogram (often due to distortion caused by VHF COMM amplifying and filtering the signal under poor communication conditions), enabling the network to understand noise in the audio. By concatenating the textual confidence vector feature with the noise vector, a comprehensive feature vector for the individual speech segment is created.

For variable-length audio inputs, extracting the overall audio quality from the temporal domain requires more than merely averaging the vectors of segments. Simple averaging does not effectively utilize the temporal information of the audio. Therefore, a transformer-based feature extraction module is utilized to extract audio quality in the temporal domain. The pre-trained feature extraction network not only focuses on the composite feature vectors of segments performing exceptionally well or poorly but also scores based on the overall quality of the speech signal globally, leveraging the audio's temporal data. After processing all audio segments of the test audio, an attention pool is used to pool all feature vectors, yielding the final radio check score.

4.3. Multi-Feature Module

4.3.1. Self-Supervised Model Feature Extraction

In 2020, Meta introduced the unsupervised speech pre-training model Wav2vec 2.0 [22]. Its core idea involves constructing a self-supervised training objective through vector quantization (VQ) and training by applying contrastive loss on heavily masked inputs. Wav2vec 2.0 significantly improved the performance of downstream speech tasks such as Automatic Speech Recognition (ASR), Text-to-Speech (TTS), and Voice Conversion (VC) through self-supervised learning on extensive amounts of unannotated speech data (such as Libri-light). Other research has demonstrated the robustness of the Wav2vec 2.0 model in speech quality prediction tasks. Therefore, the Wav2vec 2.0 base model was chosen for initial feature extraction [23,24]. For ablation studies, we also experimented with fine-tuned pre-trained models based on Chinese datasets and English datasets [25,26]. However, using these fine-tuned models resulted in a less accurate fit of the entire network to the data. We utilized the Wav2vec 2.0 model without pre-training to address this issue. Assuming the input speech signal consists of n samples, denoted as $X = [x_1, x_2, \ldots, x_n]$, following the

methodology in [22], the input speech signal is segmented with a stride of S, at this time, the audio can be defined as Equation (1):

$$X_{seg} = [(x_1, x_2, \ldots, x_s), (x_{s+1}, x_{s+2}, \ldots, x_{2s}), \ldots, (x_{n-s+1}, x_{n-s+2}, \ldots, x_n)] \quad (1)$$

After segmenting and encoding each segment, we denote the encoding as $f(\cdot)$ and the resulting encoded output as Q, defined as Equations (2) and (3).

$$Q = f(X_{seg}) = [Q_1, Q_2, \ldots, Q_m] \quad (2)$$

$$Q_i = [q_{i1}, q_{i2}, \ldots, q_{iM}] \quad (3)$$

where m represents the number of segments after segmentation, and M represents the dimensionality of the audio mapped through wav2vec 2.0 for each segment.

After this, the generated features will enter a four-layer deep feedforward network for feature transformation. The workflow can be defined as Equations (4) and (5). Assume the transformed vector is denoted as $H_{\mathrm{ssl}}^{\mathrm{k}}$, and k represents which layer of the linear transformation it has passed through.

$$H_{ssl}{}^{0} = BN(Q) \quad (4)$$

$$H_{ssl}{}^{k} = \mathrm{Re}LU(BN(Linear(H_{ssl}{}^{k-1})))(k = 1, 2, 3, 4) \quad (5)$$

where $BN()$ represents batch normalization, $Linear()$ denotes the linear layer, and $\mathrm{Re}LU()$ stands for rectified linear unit. In the subsequent ablation experiments, it will be demonstrated that this four-layer linear transformation contributes to improving the network's performance.

4.3.2. Mel Spectrogram Feature Extraction

It is common practice to transform audio into a spectrogram in order to better comprehend its frequency distribution when working with audio. the mel spectrogram is a specialized type of spectrogram that uses mel filters to map spectral data to a frequency scale that is more compatible with human auditory perception. This transformation helps us accurately capture variations in tone and timbre in the audio, as human perception of frequency is nonlinear. Therefore, this feature can effectively reflect the perception of noise in the radiotelephone communication by the air traffic controller. Moreover, studies such as [13] have demonstrated that mel spectrograms are robust in predicting speech quality.

Similarly, here, we assume that the input audio signal consists of n sample points, denoted as $\mathrm{X} = [x_1, x_2, \ldots, x_n]$. Following the computation process for mel spectrograms, the signal is segmented with a step size of S as follows at this time, the audio can be defined as Equation (6):

$$X_{seg} = [(x_1, x_2, \ldots, x_s), (x_{s+1}, x_{s+2}, \ldots, x_{2s}), \ldots, (x_{n-s+1}, x_{n-s+2}, \ldots, x_n)] \quad (6)$$

For each segment, a discrete Fourier transform (DFT) is performed, followed by mel filtering using a set of M mel filters. The result is subjected to a logarithmic transformation to obtain the mel spectrogram C defined as Equations (7) and (8).

$$C = [C_1, C_2, \ldots, C_M] \quad (7)$$

$$C_i = \log(\textstyle\sum_{i=1}^{M} H_k \left| DFT(X_{seg}) \right|^2) \quad (8)$$

In the mel filtering process, the center frequency of each filter is denoted as $Mel(k)$.H_k represents the gain of the k-th mel filter at the i-th frequency point, determining the energy contribution of that frequency point on the mel frequency scale.

After this, the generated features enter a six-layer convolutional neural network (CNN) for feature transformation, The workflow can be defined as Equations (9) and (10), where

H^k_{mel} represents the transformed vector and k indicates the layer number that has passed through the convolutional layer.

$$H^0_{mel} = C \tag{9}$$

$$H^k_{mel} = \text{Re}LU(BN(Conv2D(H^{k-1}_{mel})))(k = 1, 2, \ldots, 6) \tag{10}$$

where $BN()$ represents batch normalization, $Conv2D()$ stands for a two-dimensional convolutional layer, and $\text{Re}LU()$ represents the rectified linear unit activation function. Subsequent ablation experiments will demonstrate that this six-layer convolutional structure facilitates the model's convergence.

Mel spectrogram features and self-supervised features are aligned in the time dimension because they are segmented into the same time steps. After being flattened, the two vectors are concatenated along the time dimension as shown in Equations (11) and (12):

$$H_{ssl} = flatten(H^4_{ssl}), H_{mel} = flatten(H^6_{mel}) \tag{11}$$

$$H = concat[H_{ssl}, H_{mel}] \tag{12}$$

4.4. Feature Extraction

A temporal network is modeled using the transformer architecture introduced in [27] to process the multi-feature fusion vectors generated in the previous step. Due to its capacity to capture global contextual information and parallelization capabilities [28–31], the transformer model is widely employed in the natural language processing and computer vision domains. Therefore, we employ the transformer model as one of the modules for extracting temporal information from the network.

Firstly, the input sequence of the given data is embedded into a n-dimensional vector space. Next, position encoders are utilized to keep track of the sequence's order and relative positions of each component. These positional encodings are concatenated with the original vectors of dimension and fed into the encoder. Two major components make up the encoder: the multi-head attention (MHA) mechanism and the feedforward layer.

The MHA mechanism consists of multiple scaled dot-product attention units. Given a sequence vector, attention units calculate contextual information about specific tokens and combine them with weighted combinations of similar tokens. During training, the attention units learn three weight matrices: the key weight matrix W_K, the value weight matrix W_V, and the query weight matrix W_Q. Finally, attention representations for all tokens are obtained. The workflow can be defined as Equations (13) and (14).

$$K = HW_K, V = HW_V, \text{Q} = HW_Q \tag{13}$$

$$W_{attention} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{14}$$

where K^T represents the transpose of matrix K, while $\sqrt{\text{d}}$ denotes the dimensionality of vector features, introduced to stabilize the gradient. The softmax() function is applied for weight normalization. During computation, the weights are dot-multiplied with the attention vectors of each head, and the result is aggregated through a layer consisting of a linear transformation activated by the gelu() function. The workflow can be defined as shown in Equation (15).

$$\text{X}_{\text{attention}} = Linear\left(\text{concat}\left[W^1_{attention} \cdot H, W^2_{attention} \cdot H, \ldots, W^8_{attention} \cdot H\right]\right) \tag{15}$$

Here, $W^k_{attention}$ represents the attention mechanism generated by the k-th head, while $concat[]$ denotes the concatenation operation.

In experiments conducted by researchers such as [13,24,32], attention pooling has been shown to be effective. Therefore, we employ attention pooling at the final stage of the network to generate comprehensive quality for each segment. By averaging these

values, we obtain the ultimate prediction for speech quality. The workflow can be defined as follows:

$$\hat{y} = \frac{1}{m}\sum_{i=1}^{m} \text{softmax}(\text{Mask}(WX_{\text{attention}}))^{T} X_{\text{attention}} \quad (16)$$

where m represents the m speech segments generated in the first step, and W signifies the linear weights of the attention pool. $Mask()$ refers to the masking function that conceals vectors greater than N. Assuming the input variable is X_N^{input}, the specific procedure is as follows (17):

$$X_i^{\text{Mask}} = \begin{cases} X_i^{\text{input}}, i \leq N \\ 0, i > N \end{cases} \quad (17)$$

5. Experimental Settings

This study utilized two datasets: the NISQA dataset and the RadioCheckSpeech dataset. The NISQA dataset contains more than 14,000 speech samples, covering simulated conditions (such as effects of encoder-decoder, packet loss, and ambient noise) as well as real-world scenarios (including mobile, Zoom, Skype, and WhatsApp communications). Each audio file in this dataset is annotated with subjective ratings for individual speech segments.

This research aims to automate radio check scoring using well-established theoretical models for mean opinion score (MOS) prediction. Therefore, prior to radio check score prediction, the same model was employed for corresponding MOS score prediction to validate the model's effectiveness. As perceived differences between MOS and radio check scores were identified in Chapter 2, addressing the disparity required the use of distinct training sets for model training. Accordingly, training and validation for MOS score prediction were conducted with the NISQA_LIVE and NISQA_SIM subsets of the publicly available NISQA dataset, using the same procedures as NISQA. Additionally, for radio check score prediction, 80% of the RadioCheckSpeech dataset was used for training, while 20% was reserved for validation, as depicted in the following Figure 7.

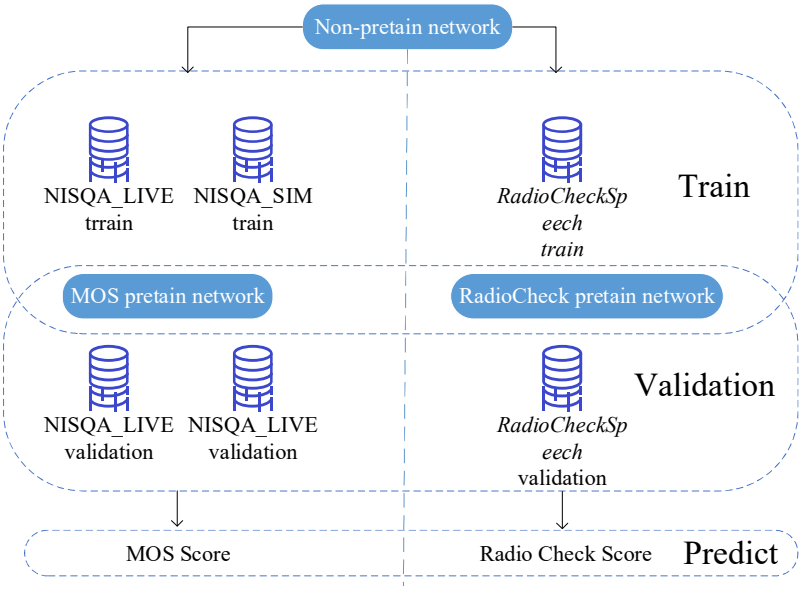


Figure 7. Methods for network training, validation, and prediction.

During the training phase, the Adam optimizer was used, which combines the advantages of the AdaGrad and RMSProp optimizers and enables rapid convergence of the model to optimal results. The Adam optimizer was initialized with a 0.0001 learning rate. The root mean square error (RMSE) was adopted as the loss function for the experiments. The network architecture was implemented using the Python programming language's

PyTorch library. The experimental computations were conducted on an RTX 4090 GPU platform. The model hyperparameters are as shown in Table 2. In the experiments, the depth of the convolutional neural network (CNN) and the deep feedforward network was selected to be four and six layers, respectively. This choice is based on the observation that networks that are either too shallow or too deep can lead to degraded performance.

Table 2. ARCnet network hyperparameter settings.

Name	Settings
Self-supervised model feature extraction settings	
Output dimensions	768
Deep feedforward extraction setting	
Deep	4
Output dimensions	384
Mel spectrogram feature extraction settings	
Number of mel bands	48
DFT window length	0.02
Maximum frequency	20,000
Convolutional neural network settings	
Conv2D core size	3×3
Conv2D dropout	0.2
Conv2D deep	6
Self-attention network settings	
Number of attention heads	8
Attention network dimension	64
Self-attention depth	2
Number of hidden units	64
Dropout	0.1
Attention pool setting	
Number of hidden units	128
Train settings	
Max train epochs	500
Train learning rate	0.0001
Train batch size	4
Train early stop epochs	20

6. Results

6.1. Quantitative Result

To verify the generalization of our network across diverse datasets, we employed standard evaluation metrics, including the Pearson correlation coefficient (r) and root mean square error (RMSE), to quantify the disparities between predicted values and actual values. The calculation formulas, Equations (18)–(20), are shown as follows:

$$\overline{\mathrm{y}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y_i}, \overline{\hat{\mathrm{y}}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \hat{\mathrm{y}}_\mathrm{i} \tag{18}$$

$$r = \frac{\sum_{i=1}^{n} (y_\mathrm{i} - \overline{y})(\hat{y}_\mathrm{i} - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_\mathrm{i} - \overline{y})}\sqrt{\sum_{i=1}^{n} (\hat{y}_\mathrm{i} - \overline{\hat{y}})}} \tag{19}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} [\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}]^2} \tag{20}$$

where $y(i)$ represents the ground truth values and $\hat{y}(i)$ denotes the predicted values provided by the network.

Similarly, we compared our method to other publicly available audio quality estimation algorithms, including P.563 [14], ANIQUE+ [15], WAWEnets [16], and dual-end models POLQA [7], DIAL [17], and VISQOL (v3.1.0) [18]. As our baseline model, we selected the NISQA model, which demonstrated superior performance.

Table 3 shows the test results of all models on the NISQA open-source dataset. Compared with other algorithms, our model achieves the best performance, and compared to the baseline model NISQA, our model achieves a relative performance improvement of 6% in Pearson correlation coefficient and 12% in RMSE on the NISQA open-source dataset, demonstrating our model's effectiveness. On the RadioCheckSpeech dataset, our model achieves a relative performance improvement of 12% in Pearson correlation coefficient and 12% in RMSE, indicating that our model achieves the best results for both MOS score prediction and radio check score prediction tasks. Due to the fact that the RadioCheckSpeech dataset does not contain clean speech, the bi-end model does not provide test results for the RadioCheckSpeech dataset.

Table 3. Performance of the model on the NISQA and RadioCheckSpeech datasets.

Dataset	**NISQA_VAL_LIVE**		**NISQA_VAL_SIM**		**RadioCheckSpeech**	
Files	**200**		**2500**		**641**	
Model	**r↑**	**RMSE↓**	**r↑**	**RMSE↓**	**r↑**	**RMSE↓**
ARCnet	0.88	0.39	0.92	0.44	0.93	0.43
NISQA	0.83	0.39	0.89	0.50	0.83	0.49
P503	0.45	0.52	0.44	1.00	0.42	0.82
ANIQUE+	0.52	0.59	0.53	0.96	0.54	0.65
WAWEnets	0.35	0.65	0.29	1.04	0.33	0.99
POLQA	0.67	0.58	0.74	0.75	\	\
DIAL	−0.11	0.73	0.35	1.24	\	\
VISQOL	0.55	0.42	0.74	0.52	\	\

6.2. Ablation Study

We conducted an ablation study to determine whether the use of different networks to preprocess the artificial features and the self-supervised features and the use of different pre-trained self-supervised models would affect the network's perceptual performance on speech quality.

Table 4 displays the results of concatenating the two features and then extracting the features using a single network (DFFnet (deep feedforward network) and CNN (convolutional neural networks)). Compared to the parameter-heavy CNN extraction network, the results indicate that a simple DFFnet network can improve the network's performance. However, it is also apparent that using different extraction methods for different features can significantly improve the model's feature extraction effect.

Table 4. Performance of the frameworks (trained with deep feedforward neural network, convolutional neural network, and their combination) on the NISQA and RadioCheckSpeech datasets.

Dataset	**ARCnet**		**DFFnet**		**CNN**	
	r↑	**RMSE↓**	**r↑**	**RMSE↓**	**r↑**	**RMSE↓**
NISQA_VAL_LIVE	0.88	0.39	0.84	0.55	0.72	0.49
NISQA_VAL_SIM	0.9	0.44	0.82	0.5	0.67	0.82
RadioCheck Speech	0.93	0.43	0.64	0.67	0.73	0.60

Table 5 shows the test results of using different self-supervised pre-trained models under the condition that other parameters are unchanged. Clearly, the use of trained pre-trained models will result in a significant decrease in network performance. A possible cause is that the self-supervised models have not been pre-trained using actual radiotelephone communications speech data. Fine-tuning the self-supervised models with radiotelephone communication speech data is also one of the future research directions to improve the network's perceptual ability.

Table 5. Performance of the frameworks (utilizing self-supervised learning models pre-trained with Chinese and English) on the NISQA and RadioCheckSpeech datasets.

Dataset	ARCnet		Chinese_Pretrain_wav2vec2.0		Pretrain_wav2vec2.0	
	r↑	RMSE↓	r↑	RMSE↓	r↑	RMSE↓
NISQA_VAL_LIVE	0.88	0.39	0.4	0.65	0.72	0.49
NISQA_VAL_SIM	0.92	0.44	0.54	0.93	0.68	0.81
RadioCheck Speech	0.93	0.43	0.73	0.60	0.74	0.59

6.3. Result Interpretability Analysis

In the discussions of the previous two sections, we described how our research enhances the performance of the ARCnet network. However, the interpretability of the ARCnet network's predictive results was not addressed. Therefore, this section has designed multiple experiments that involve adding different types of impairments to audio to explore whether the ARCnet network's perception of various types of damage (noise level, voice volume, and frequency band loss) has perceptual interpretability, thereby conducting an interpretability analysis of the ARCnet network.

Firstly, the impact of noise level on voice quality is very important. Therefore, we selected a clear control command (radio check value of 5) and added noise impairment. By adjusting the signal-to-noise ratio (SNR), we explored the acoustic interpretability of the ARCnet network in the presence of noise damage. The results are shown in Figure 8.

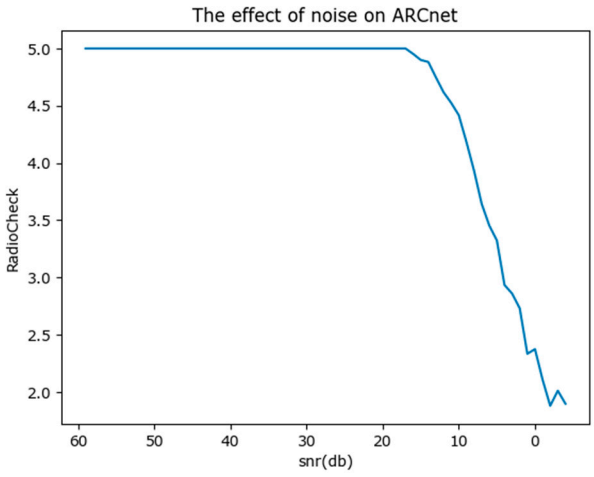


Figure 8. The effect of noise on ARCnet.

As illustrated, the impact of noise on the ARCnet network's evaluations is significant. Moreover, the radio check value sharply declines when the signal-to-noise ratio (SNR) falls below 20 dB. This is similar to the control operators' evaluation of noise damage, demonstrating that the ARCnet network's performance in the dimension of noise damage aligns with human auditory perception. The ARCnet network has acoustic interpretability in the dimension of noise damage.

Secondly, the volume of speech also directly affects the quality of the voice. Too low a volume can result in control operators not being able to hear the commands clearly, while too high a volume can cause "clipping" effects. Therefore, we also selected a segment of control speech with a clear and appropriate volume (radio check value of 5). By reducing the volume, and then increasing and clipping it (simulating microphone filtering), we introduced volume impairment. We explored the acoustic interpretability of the ARCnet network in the context of volume damage. The results are shown in Figure 9a,b.

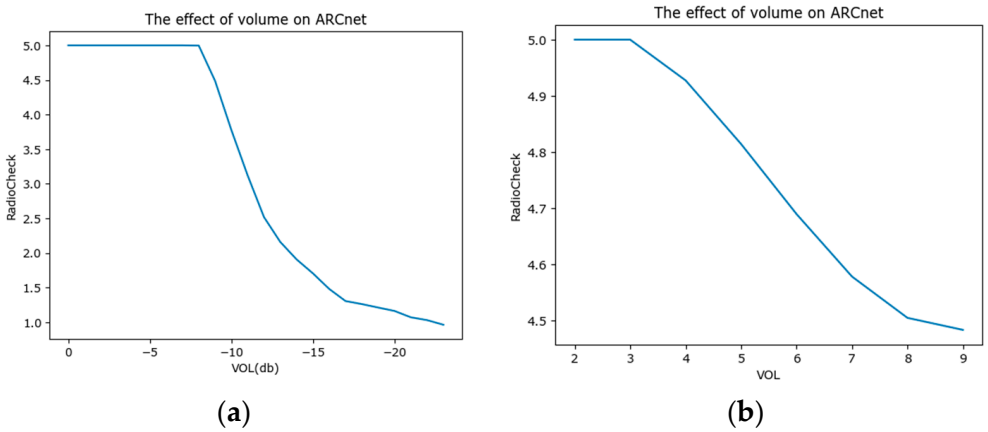


Figure 9. The effect of volume on ARCnet. (**a**) Shows the change in ARCnet scores as the audio volume decreases (dB). (**b**) Shows the change in ARCnet scores as the audio volume increases.

As shown in Figure 9a, as the volume decreases, the radio check value also declines. In fact, the perception pattern of volume impairment by control operators is similar. However, the attenuation turning point usually occurs earlier because multiple controllers in the control room will be using voice communications simultaneously, and the auditory threshold is consequently reached sooner. This also indicates that there is room for improvement in the ARCnet network's evaluation of too-low volumes.

As shown in Figure 9b, as the sound amplification increases, the voice gradually exhibits clipping effects, and the corresponding radio check scores begin to decline, which matches the perception of the controllers perfectly.

In summary, the ARCnet network also has acoustic interpretability in the dimension of volume damage.

Finally, as mentioned in the Preliminary Analysis section, compared to cellular mobile communication networks based on packet transmission, VHF COMM introduces a unique type of audio loss characterized by band-like noise in the audio spectrum curve, known as frequency band damage. Unlike traditional noise damage, frequency band damage is introduced by equipment and is characterized by concentrated energy, focused frequency, and constant duration. The masking effect of this type of damage is stronger than that of general noise damage. Additionally, the human ear perceives different results for damage occurring in different frequency bands. Therefore, to verify whether the ARCnet network's perception of frequency band damage is acoustically interpretable and similar to human ear perception, the experiment also initially selected a segment of clear control speech without band-like noise (radio check value of 5). By introducing a constant 40 dB frequency band noise in specified frequency bands to simulate frequency band damage, we explored the acoustic interpretability of the ARCnet network in the context of frequency band damage. The results are shown in Figure 10.

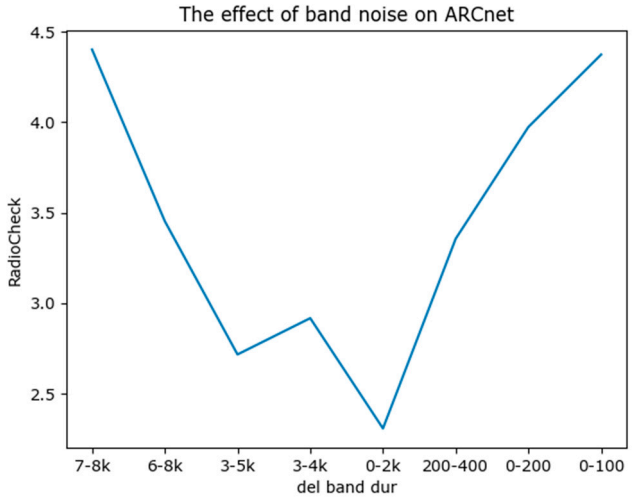


Figure 10. The effect of band noise on ARCnet.

As illustrated, the ARCnet network is particularly sensitive to mid-frequency band damage (loss concentrated between 1000 and 4000 Hz) and relatively insensitive to low or high-frequency band damage. Moreover, wider band damage has a smaller impact on ARCnet network evaluations compared to narrow band damage. Since human speech frequencies are concentrated around 1–3 kHz, damage within this frequency range noticeably affects speech, making these evaluation results similar to actual human ear perception. This still indicates that the ARCnet network's outcomes in the dimension of frequency band damage possess acoustic interpretability. Although this evaluation method is similar to human hearing, it also shows that the ARCnet network's evaluation of frequency band damage is biased and may overlook some damages in low or high frequencies, which might not be objective enough in certain evaluation settings.

In conclusion, as described in the experiments mentioned previously, the results of the ARCnet network's evaluation possess acoustic interpretability across three dimensions: noise damage, volume damage, and frequency band damage. However, the experiments still expose some of the ARCnet network's shortcomings, such as overestimating the quality of low-volume audio in volume damage and tending to overlook damage in the lower and higher frequency bands. These are areas for future improvement to enhance the ARCnet network's evaluation capabilities.

7. Conclusions and Future Work

This study aims to achieve radio check machine scoring, and for the first time, proposes ARCnet to imitate air traffic controllers' radio check call quality prediction. This study first integrated artificial features (MFCC) and self-supervised features (wav2vec 2.0) to build a multi-feature fusion model, which is used for the machine's real-time subjective quality prediction of the call by comparing the similarities and differences between MOS score and radio check score, and drawing on the research results of MOS score. The results of the ablation experiment indicate that using different feature extraction networks based on different features can significantly boost the performance of the model. It is worth noting that even for different and fine-tuned self-supervised models, their performance in the network will suffer due to the difference between the audio scene of the pre-training dataset and the radiotelephone communication scene. Based on the validation of the model's performance on the selected MOS score dataset, we created a dataset named RadioCheck-Speech for radio check score prediction. This dataset contains 3200 real radiotelephone communication speeches that were scored by controllers using the radio check scoring criteria and the actual scoring results were recorded. We fine-tuned and tested the model on this dataset and the NISQA open-source dataset. Our research indicates that, compared to other network models, ARCnet is superior at predicting subjective call quality evaluation.

Compared with the baseline model, ARCnet achieves a relative performance improvement of 6% and 12% in the Pearson correlation coefficient, and 12% and 12% in RMSE on the NISQA dataset and RadioCheckSpeech, respectively. This demonstrates the effectiveness of the radio check score prediction ARCnet model and method for creating datasets. Our experiment still has some shortcomings, such as in normal radiotelephone communication situations, the radio check scores should be above 3, and when the score reaches below 3, it can be understood that the VHF communication system has an alarm and needs to be repaired. This inevitably results in an uneven distribution of data. In the future, we will consider adding noise and other methods to the RadioCheckSpeech dataset.

In the future, ARCnet could be integrated into intercom systems as a virtual controller responsible for the radio check after the initial contact is established between the captain and the controller. However, in actual airport operations, different controllers have different thresholds during radio checks. Customizing radio check thresholds for the model based on small sample sizes for different controllers is also an important factor that ARCnet should consider in the future. Additionally, the civil aviation industry will upgrade to air-ground data link systems represented by 5G in the future [33]. In future communication systems, voice commands will be transmitted in the form of data packets. The standards for evaluating voice will likely focus on factors such as packet loss in digital communication, leading to audio discontinuities and audio delays. How ARCnet will evaluate voice communications in the air-ground data link is another potential challenge for the future. ARCnet could also be used to evaluate noise reduction algorithms for voice, assessing the effectiveness of voice denoising. Moreover, an upgraded ARCnet could serve to judge the anthropomorphism of synthetic voices. Future research on ARCnet will focus on developing multitask learning and transfer learning, and considering the real-time aspects of ARCnet in restricted environments will also be a key research focus.

Author Contributions: Conceptualization, W.P. and Y.W.; methodology, Y.W.; software, Y.W.; validation, Y.W. and W.P.; formal analysis, Y.Z. and Y.W.; investigation, Y.W.; resources, W.P.; data curation, B.H.; writing—original draft preparation, Y.W.; writing—review and editing, W.P. and Y.W.; visualization, Y.W.; supervision, Y.W. and B.H.; project administration, Y.W.; funding acquisition, W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (U2333209), National Key R&D Program of China (No. 2021YFF0603904), National Natural Science Foundation of China (U1733203) and Safety Capacity Building Project of Civil Aviation Administration of China (TM2019-16-1/3), the Fundamental Research Funds for the Central Universities (PHD2023-044).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations and symbols are used in this manuscript:

VHF COMM	Very high frequency communication system
GMDSS	Global maritime distress and safety system
VHF	Very high frequency
AM	Amplitude modulation
ATC	Air traffic controller
ICAO	International civil aviation organization
MOS	Mean opinion score
PESQ	Perceptual evaluation of speech quality
VOIP	Voice over Internet protocol
t-SNE	t-distributed stochastic neighbor embedding
RMSE	Root mean square error
VQ	Vector quantization
ASR	Automatic speech recognition

TTS	Text-to-speech
VC	Voice conversion
DFT	Discrete Fourier transform
CNN	Convolutional neural network
MHA	Multi-head attention
DDFnet	Deep feedForward network
MFCC	Mel-frequency cepstrum coefficient
Symbols	
X	Sampled audio signal
X_{seg}	Segmented audio signal
$f(\cdot)$	Encoding method of Wav2vec 2.0
Q	Encoding result of Wav2vec 2.0
$H_{\text{ssl}}^{\text{k}}$	Output of deep feedforward network
$BN()$	Batch normalization
$Linear()$	Linear layer
$\text{Re}LU()$	Rectified linear unit
M	Number of mel filters
C	Mel spectrogram
$DFT()$	Discrete Fourier transform
$Mel()$	Mel filtering process
$\log()$	Logarithm
H_{mel}^{k}	Output of convolutional neural network
$Conv2D()$	Two-dimensional convolution
$flatten()$	Flatten
W_K	Key weight matrix
W_V	Value weight matrix
W_Q	Query weight matrix
K^T	Transpose of matrix K
$\sqrt{\text{d}}$	Dimensionality of vector features
$softmax()$	Softmax normalization
$gelu()$	Gaussian error linear unit
$W_{attention}$	Output of attention mechanism
$concat[]$	Concatenation operation
$Mask()$	Masking function
y	Ground truth values
$\hat{y}$	Predicted values
r	Pearson correlation coefficient
$RMSE$	Root mean square error

References

1. ICAO, D. 4444 ATM/501. Procedures for Air Navigation Services Traffic Management. 2001. Available online: https://www.ealts.com/documents/ICAO%20Doc%204444%20Air%20Traffic%20Management.pdf (accessed on 6 May 2024).
2. Doc, I. 9432 An/925. Manual of Radiotelephony. 2007. Available online: https://www.ealts.com/documents/ICAO%20Doc%209432%20Manual%20of%20Radiotelephony%20(4th%20ed.%202007).pdf (accessed on 6 May 2024).
3. Doc, I. 9870 AN/463. Manual on the Prevention of Runway Incursions. 2006. Available online: https://www.icao.int/safety/RunwaySafety/Documents%20and%20Toolkits/ICAO_manual_prev_RI.pdf (accessed on 6 May 2024).
4. Shattil, S.; Alagar, A.; Wu, Z.; Nassar, C. Wireless Communication System Design for Airport Surface Management .1. Airport Ramp Measurements at 5.8 GHz. In Proceedings of the 2000 IEEE International Conference on Communications. ICC 2000. Global Convergence through Communications. Conference Record, New Orleans, LA, USA, 18–22 June 2000; Volume 3, pp. 1552–1557.
5. Pinska-Chauvin, E.; Helmke, H.; Dokic, J.; Hartikainen, P.; Ohneiser, O.; Lasheras, R.G. Ensuring Safety for Artificial-Intelligence-Based Automatic Speech Recognition in Air Traffic Control Environment. *Aerospace* **2023**, *10*, 941. [CrossRef]
6. P. 23, I.-T.S. Coded-Speech Database. 1998. Available online: https://www.itu.int/rec/T-REC-P.Sup23-199802-I (accessed on 6 May 2024).
7. Assessment, P.O.L.Q. Document ITU-T Rec. P 863. Int. Telecommun. Union. 2011. Available online: https://www.itu.int/rec/T-REC-P.863 (accessed on 6 May 2024).

8. Malfait, L.; Berger, J.; Kastner, M.P. 563—The ITU-T Standard for Single-Ended Speech Quality Assessment. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1924–1934. [CrossRef]
9. Lo, C.-C.; Fu, S.-W.; Huang, W.-C.; Wang, X.; Yamagishi, J.; Tsao, Y.; Wang, H.-M. Mosnet: Deep Learning Based Objective Assessment for Voice Conversion. *arXiv* **2019**, arXiv:1904.08352.
10. Fu, S.-W.; Tsao, Y.; Hwang, H.-T.; Wang, H.-M. Quality-Net: An End-to-End Non-Intrusive Speech Quality Assessment Model Based on BLSTM. *arXiv* **2018**, arXiv:1808.05344.
11. Yoshimura, T.; Henter, G.E.; Watts, O.; Wester, M.; Yamagishi, J.; Tokuda, K. A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 342–346.
12. Patton, B.; Agiomyrgiannakis, Y.; Terry, M.; Wilson, K.; Saurous, R.A.; Sculley, D. AutoMOS: Learning a Non-Intrusive Assessor of Naturalness-of-Speech. *arXiv* **2016**, arXiv:1611.09207.
13. Mittag, G.; Naderi, B.; Chehadi, A.; Möller, S. Nisqa: A Deep Cnn-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. *arXiv* **2021**, arXiv:2104.09494.
14. Rec, I.P. *563: Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*; International Telecommunication Union: Geneva, Switzerland, 2004; pp. 1–25.
15. Kim, D.-S.; Tarraf, A. ANIQUE+: A New American National Standard for Non-Intrusive Estimation of Narrowband Speech Quality. *Bell Labs Tech. J.* **2007**, *12*, 221–236. [CrossRef]
16. Catellier, A.A.; Voran, S.D. Wawenets: A No-Reference Convolutional Waveform-Based Approach to Estimating Narrowband and Wideband Speech Quality. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 331–335.
17. Côté, N. *Integral and Diagnostic Intrusive Prediction of Speech Quality*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
18. Chinen, M.; Lim, F.S.; Skoglund, J.; Gureev, N.; O'Gorman, F.; Hines, A. ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric. In Proceedings of the 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, 26–28 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
19. Ragano, A.; Skoglund, J.; Hines, A. NOMAD: Unsupervised Learning of Perceptual Embeddings for Speech Enhancement and Non-Matching Reference Audio Quality Assessment. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1011–1015.
20. Fu, S.-W.; Hung, K.-H.; Tsao, Y.; Wang, Y.-C.F. Self-Supervised Speech Quality Estimation and Enhancement Using Only Clean Speech. *arXiv* **2024**, arXiv:2402.16321.
21. Liu, H.; Liu, M.; Wang, J.; Xie, X.; Yang, L. Non-Intrusive Speech Quality Assessment with Multi-Task Learning Based on Tensor Network. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 851–855.
22. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
23. Tseng, W.-C.; Huang, C.; Kao, W.-T.; Lin, Y.Y.; Lee, H. Utilizing Self-Supervised Representations for MOS Prediction. *arXiv* **2021**, arXiv:2104.03017.
24. Oliveira, F.S.; Casanova, E.; Júnior, A.C.; Gris, L.R.S.; Soares, A.S.; Galvão Filho, A.R. Evaluation of Speech Representations for MOS Prediction. In Proceedings of the International Conference on Text, Speech, and Dialogue, Pilsen, Czech Republic, 4–6 September 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 270–282.
25. Zhang, B.; Lv, H.; Guo, P.; Shao, Q.; Yang, C.; Xie, L.; Xu, X.; Bu, H.; Chen, X.; Zeng, C.; et al. Wenetspeech: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6182–6186.
26. Hsu, W.-N.; Sriram, A.; Baevski, A.; Likhomanenko, T.; Xu, Q.; Pratap, V.; Kahn, J.; Lee, A.; Collobert, R.; Synnaeve, G.; et al. Robust Wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. *arXiv* **2021**, arXiv:2104.01027.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
28. Yu, J.; Li, J.; Yu, Z.; Huang, Q. Multimodal Transformer with Multi-View Visual Representation for Image Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 4467–4480. [CrossRef]
29. Shi, Y.; Wang, Y.; Wu, C.; Fuegen, C.; Zhang, F.; Le, D.; Yeh, C.-F.; Seltzer, M.L. Weak-Attention Suppression for Transformer Based Speech Recognition. *arXiv* **2020**, arXiv:2005.09137.
30. Wu, C.; Xiu, Z.; Shi, Y.; Kalinli, O.; Fuegen, C.; Koehler, T.; He, Q. Transformer-Based Acoustic Modeling for Streaming Speech Synthesis. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 146–150.
31. Zhang, R.; Wu, H.; Li, W.; Jiang, D.; Zou, W.; Li, X. Transformer Based Unsupervised Pre-Training for Acoustic Representation Learning. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 6933–6937.

32. Safari, P.; India, M.; Hernando, J. Self-Attention Encoding and Pooling for Speaker Recognition. *arXiv* **2020**, arXiv:2008.01077.
33. Iqbal, S.; Hamamreh, J.M. A Comprehensive Tutorial on How to Practically Build and Deploy 5G Networks Using Open-Source Software and General-Purpose, off-the-Shelf Hardware. *RS Open J. Innov. Commun. Technol.* **2021**, *2*, 1–28. [CrossRef]