SpeechQoE: A Novel Personalized QoE Assessment Model for Voice Services via Speech Sensing

Chaowei Wang
chaowei.wang@mavs.uta.edu
The University of Texas at Arlington

Huadi Zhu
huadi.zhu@mavs.uta.edu
The University of Texas at Arlington

Ming Li
ming.li@uta.edu
The University of Texas at Arlington

ABSTRACT

Quality of Experience (QoE) assessment is a long-lasting but yet-to-be-resolved task. Existing approaches, especially for conversational voice services, are restricted to leveraging network-centric parameters. However, their performances are hardly satisfactory due to the failure to consider comprehensive QoE-related factors. Moreover, they develop a *one-for-all* model that is uniform for all individuals and thus incapable of handling user diversity in QoE perception. This paper proposes a personalized QoE assessment model, namely SpeechQoE. It exploits speaker's speech signals to infer individual's perceived quality in voice services. SpeechQoE fundamentally addresses the drawback of conventional models. Instead of enumerating and incorporating unlimited QoE-related factors, SpeechQoE takes as input speech signals that inherently bear rich information needed for QoE assessment of the speaker. SpeechQoE employs an efficient few-shot learning framework to adapt the model to a new user quickly. We additionally design a lightweight data synthetic scheme to minimize the overhead of data collection needed for model adaption. A modular integration with a conventional parametric model is further implemented to avoid issues caused by the clean-slate data-driven approach. Our experiments show that SpeechQoE achieves an accuracy of 91.4% in QoE assessment which outperforms the state-of-the-art solutions by a clear margin. As another contribution of this work, we build a dataset that would be the first source of annotated audio tracks for QoE assessment of conversational calls.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**; • **Computing methodologies** → **Machine learning algorithms**.

ACM Reference Format:
Chaowei Wang, Huadi Zhu, and Ming Li. 2022. SpeechQoE: A Novel Personalized QoE Assessment Model for Voice Services via Speech Sensing. In *The 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22), November 6–9, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3560905.3568502

1 INTRODUCTION

Motivation. During the pandemic, shelter-in-place, work-from-home, school-from-home, and other new hybrid lifestyles have been adopted globally, either mandated by local ordinances or voluntarily recommended by companies. Under the "new normal", various audio/video telecommunication apps, such as Skype, FaceTime, Zoom, and Microsoft Teams, play an irreplaceable role that enables people to communicate in real time with geographically dispersed peers via terminal devices. Many of them have been reported with recorded surges from every aspect, ranging from the usage time to the number of registered customers. In the meantime, it is always the center of interest for service providers to get an in-depth understanding of user's satisfactory levels of service quality.

Extensive prior efforts have been devoted to QoE modeling. They aim to map a diverse spectrum of impact factors to a QoE score given a specific multimedia service type. The mainstream approaches develop the so-called *parametric models* [8, 12, 14–16, 19, 30, 70]. They estimate user's QoE through the characterisation of underlying networks, e.g., source rate, packet loss, coding scheme, delay, and jitter. In fact, QoE is affected by many other factors, which are usually grouped by human factors (e.g., mood, expectation) and contextual factors (e.g., background noise level). In other words, conventional wisdom fails to capture a full spectrum of QoE-related factors. Besides, parametric models are uniform for all individuals. Instead, people's quality perception is highly subjective and heterogeneous. For example, some users are more sensitive to echoing sound in a conversational call, while others care more about latency. A *one for all* model inevitably performs poorly. In summary, a new perspective on QoE assessment is needed to attain a fundamentally better performance. Our discussion pertains to voice services, i.e., the audio conversational calls carried over either telephony networks or various telecommunication apps. We will leave QoE modeling for the video calls in our future work.

Our approach. Our idea is inspired by a well-recognized phenomenon in neurophysiology that characteristics of a speaker's speech can reflect her subjective feelings [21, 25, 47, 72]. For example, speech produced in a state of fear, anger, or joy becomes loud and fast, with a higher and broader range in pitch, whereas emotions such as sadness or tiredness generate slow and low-pitched speech [51, 64]. Motivated by the observation, we seek to answer a key question: Can we design a system to leverage the novel speech indicator to assess QoE? To this end, we propose SpeechQoE, a personalized QoE assessment model for voice services via speech sensing.

The model takes the mic recorded speech as inputs and produces a corresponding QoE score. The ITU-recommended five-level MOS score is adopted, where 5 represents the best quality and 1 represents the worst. QoE assessment is then cast into a classification problem. To extract explicit and implicit features from speech signals, we employ a convolutional neural network (CNN). Note that deriving a uniform model via the data-driven approach

SenSys '22, November 6–9, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9886-2/22/11.
https://doi.org/10.1145/3560905.3568502

is trivial. Rather, we are interested in constructing a personalized model; that is, parameters in the CNN classifier are user-specific. The model personalization is achieved by fine-tuning the CNN classifier, pre-trained offline via a general training dataset, to adapt to each individual. Specifically, *few-shot learning* technique [27, 28] is applied for quick model adaption in few shots, e.g., 3-5 annotated samples of each class from the new user. In a holistic view, SpeechQoE is superior to the state-of-the-art approaches in three aspects. First, it exploits subjective factors, i.e., speech-based cues, for QoE assessment; they have been largely overlooked by prior works. Second, existing models endeavor to exhaustively enumerate and incorporate various impact factors in QoE modeling, which are impractical to implement in real-world scenarios. Alternatively, our approach merely utilizes speech to reflect the speaker's perceived QoE as a whole. Most importantly, SpeechQoE is a personalized model that is fine-tuned to each individual to deliver the optimized assessment accuracy.

Despite the nice properties of SpeechQoE, its implementation is faced with a critical challenge. Under the framework of few-shot learning, the model fine-tuning requires the existence of several labeled samples of each class (i.e., QoE levels) from the new user. However, network conditions nowadays do not deteriorate often that produce very low QoE scores, especially for voice services that consume relatively narrow bandwidth. Therefore, it may take a considerable amount of time, say, several weeks, to gather sufficient samples; otherwise, the model personalization cannot be successfully performed. To resolve this issue, we enhance the few-shot learning framework with a lightweight data synthesis scheme that augments the new user's dataset. To be specific, we first identify the new user's "close neighbors", which share high similarity in data distribution with the new user, from the existing training dataset. We then calibrate the new user's data distribution through those from her close neighbors. An adequate number of samples, covering all QoE scores, are sampled from the calibrated distribution. We then fine-tune the model via a joint set of real and synthetic data samples. In this way, the data collection overhead can be reduced significantly. Compared with conventional data synthesis and data augmentation methods [50, 53, 63, 69, 77], our approach can deal with the absence of samples from certain QoE level(s), a unique situation in our problem.

The proposed sample synthesis method relies on an adequate number of close neighbors. In practice, this assumption may not always hold when the user pool is not large enough. As a last piece of the jigsaw puzzle, our design is integrated with the parametric model. It serves as a protective backup and is activated once insufficient close neighbors are identified. The hybrid design prevents the QoE assessment from catastrophic performance degradation.

To evaluate SpeechQoE, we develop a prototype and conduct in-person experiments in a lab setting. Our collected speech dataset contains 190 hours of clean speech signals from 38 users. Our evaluation results show that SpeechQoE can achieve 91.4% accuracy on average. It beats three state-of-the-art baselines by 42.5%, 18.4%, and 13.0% , respectively. It delivers consistent performances over a variety of calling environments. SpeechQoE represents the first personalized QoE assessment model leveraging speech signals while overcoming the data collection issue. Through the SpeechQoE design, we make the following contributions:

- We investigate the relationship between QoE and speech signals through a measurement study.
- We introduce SpeechQoE, a personalized QoE assessment model for voice services using speech-based cues. Compared with conventional models, ours exploits subjective factors to characterize user's perceived service quality. More importantly, it is a user-specific model that takes into account user diversity in QoE modeling.
- We renovate the conventional few-shot learning framework by introducing a novel data synthesis scheme. It allows the model to quickly adapt to a new user within limited samples. The data synthesis scheme can potentially be applied to other data-hungry cases.
- We build our own dataset[1] via a six-month data collection campaign. 38 volunteers and 5 student workers get involved. To our knowledge, it would be the first data source of annotated audio tracks for conversational call QoE assessment. Extensive tests are performed over our dataset. Results validate the efficacy and efficiency of the new QoE model.

The rest of this paper is organized as follows. Section 2 covers necessary background of leveraging speech for QoE assessment. A measurement study that validates the feasibility of our idea is presented as well. A basic QoE assessment model using few-shot learning and its limitation are discussed in Section 3. In Section 4, we presents an advanced model that can quickly adapt to a new user within limited samples. We evaluate SpeechQoE in Section 5. Section 6 reviews prior works related to our topic. A discussion regarding potential future works is provided in Section 7. We conclude the paper in Section 8. The entire study is IRB-approved.

2 BACKGROUND

The object of this part is to validate the feasibility of exploiting speech patterns for QoE assessment of voice services.

2.1 Speech as A Cue for Subjective Perceptions

A strong connection between acoustic properties of speech signals and human perceptions has been accepted for a decade [21, 25, 47, 72]. Such properties include prosodic (i.e., pitch, loudness, and rhythm) and voice quality. In neurophysiological literature, it is demonstrated that phonation, respiration, and articulation in speech are unconsciously regulated by autonomic nervous system stimulation, which is known to produce responsive output under numerous emotional states. For example, speech made in a state of fear, anger, or joy becomes loud and fast, with a higher and wider range in pitch, whereas emotions such as sadness or tiredness generate slow and low-pitched speech [51, 64]. Mostly based on established procedures in phonetics and speech sciences, researchers have explored a large number of acoustic parameters, both in the time domain (e.g., speaking rate and zero-crossing rate) and the frequency domain (e.g., fundamental frequency F_0, formant frequencies, and intensity or energy in different frequency bands) to assess speaker's emotional perception. For example, Alpert *et al.* [9] found that depression leads to reduced intensity, lower speaking rate, and narrower pitch range. Gudmalwar *et al.* [34] observed that zero-crossing rate (ZCR) is higher in arousal emotion states (e.g.,

[1]The dataset is open-sourced via https://github.com/MobiSec-CSE-UTA/SpeechQoE

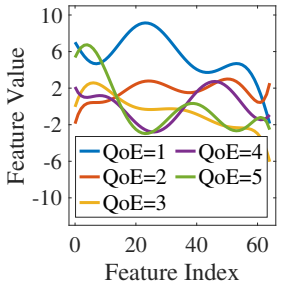


Figure 1: Feature vectors associated with different QoE scores.

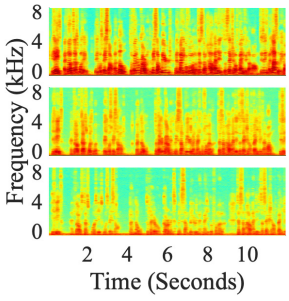


Figure 2: Top-to-down: Spectrogram of audio segments with QoE scores 1, 3 and 5.

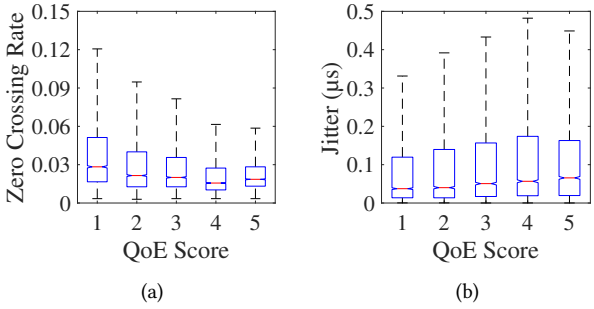


Figure 3: Comparison of statistics of speech features under different QoE perceptions. (a) Zero crossing rate (ZCR). (b) Jitter.

anger) compared with suppressed emotion states (e.g., sadness). Hammerschmidt and Jurgens [35] performed acoustic analysis on a set of speech utterances. The findings confirm that acoustic parameters extracted from both time and frequency domains are important cues of subjective feelings. All these findings motivate us to investigate the feasibility of leveraging speech-based cues to infer subjective QoE perception on voice services.

2.2 Measurement Study

The measurement study intends to explore potential correlations between speech-based cues and QoE perception.

Measurement setup. Six subjects are involved in the measurement study. Two form a pair and conduct calling sessions in two separate rooms via a VoIP platform we set up. We simulate calling environments by varying network conditions in the platform and background noise in rooms to render voice services of various qualities. After each calling session, subjects are asked to rate the service quality from 1 to 5, with 1 the worst and 5 the best. We record at least 8 calls from each subject under each QoE score. More details about experiment setup are elaborated in Section 5.1.

Observation 1: Speech patterns correlate with perceived QoE. We extract a standard acoustic parameter set, Geneva minimalistic acoustic parameter set (GeMAPS) [23], which consists of 65 features, such as pitch, jitter, and MFCC etc., out of recorded audio tracks from the participants. OpenSMILE [24] is used, an open-source framework for feature extraction from audio signals. For Figure 1, the data is from one randomly selected subject out of six. All the features are extracted from 40 recorded audio tracks from the subject. There are 8 tracks under each QoE score. The feature values are the average result over 8 tracks. The curves are obtained by further applying a Polynomial regression over the 65 features. We observe in Figure 1 that features exhibit high inter-class difference in their values. We further examine the spectrogram of three audio segments labeled with different QoE scores in Figure 2. To avoid the impact from individual heterogeneity, all speech segments are from the same subject. Short-time Fourier transform (STFT) is applied, with a 0.2 s length Hann window and 0.01 s length hop. It is evidenced that the spectrogram intensity is higher at a lower QoE score across the frequency components. It implies that speakers tend to raise their voices as experiencing a worse calling quality. This phenomenon meets our expectation–people speak louder to be heard by their peers under a poor conversation condition.

We also show the distribution of two acoustic features, zero crossing rate (ZCR) and jitter, under different QoE perception in Figure 3a and Figure 3b, respectively. The data is from one randomly selected subject. The two features are derived from 40 recorded audio tracks from the subject. There are 8 tracks under each QoE score. The feature values are the average result over 8 tracks. Both figures demonstrate clear differences in feature statistics with respect to QoE scores. Specifically, ZCR measures how many times the waveform crosses the zero axis. It is an indicator of the smoothness of audio waveforms. Speaker's speech is observed varying more in her sound under a worse quality perception. Jitter, different from the network parameter jitter, is quantified as the cycle-to-cycle variations of fundamental frequency. A more significant jitter usually indicates faster speaking. Figure 3b shows that people tend to speak faster when experiencing a better calling condition.

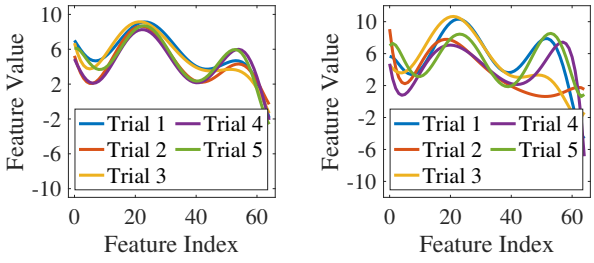


Figure 4: Comparison of feature vectors extracted from different trials (QoE=1) under (a) the same contextual speaking content and (b) different contents.

Observation 2: Speech-QoE correlation is consistent. Figure 4a compares the feature vector in five trials under the same QoE score (QoE=1). The five trials are randomly picked from all recorded audio tracks from one participant. The features exhibit high consistency. It implies that the speech-QoE correlation discussed above is not a sporadic event but a persistent property. Note that we adopt the same contextual speaking content in these trials. We are also interested in examining if the consistency would be impacted when contents are changed. Figure 4b shows the feature

vector in another five trials, each associated with a unique conversation. The sentences involved in each conversation are different. The consistency is still observed, though some variations are introduced. Besides, it is noteworthy that some features are more robust against content diversity than others. This insight suggests that those content-agnostic features should play a more significant role in our QoE assessment model.

Observation 3: Pattern diversity among users. Speech as one kind of biometric, it exhibits diverse patterns among individuals. We use pitch as an illustration. Particularly, pitch is the relative highness or lowness of a tone as perceived by the ear. It depends on the number of vibrations per second produced by the vocal cords. Figure 5a shows the pitch distribution of two randomly selected subjects from two trials under the same QoE score. The distributions are distinguishable. As a comparison, those from the same person in two trials are similar. We further show in Figure 5b a correlation matrix (15×15) on pitch values among five participants, each involving three trials. A similar observation is obtained–features extracted from speakers' speech are diverse, even though they may perceive the same quality of service. It meets our expectation: Each person's vocal tract is unique; physical features, both phonetic and morphological, are particular to each individual.

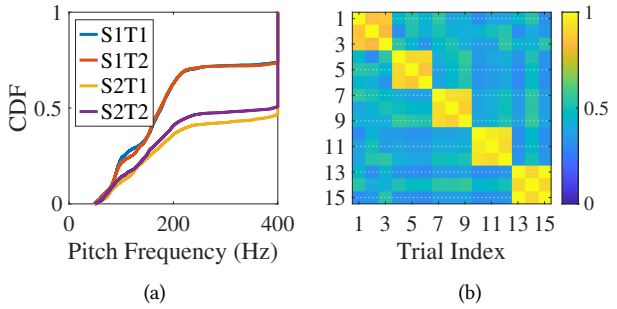


Figure 5: Illustration of speech pattern diversity among users. (a) CDF of pitch from two subjects (S) each involving two trials (T). (b) Correlation matrix among 15 tests: 5 subjects each involving 3 trials.

Summary. Our findings are encouraging. First, speaker's speech patterns are distinguishable under different QoE scores, that is, speaker's satisfactory level toward service quality can be reflected by her speech response. Second, the relationship between the speech pattern and QoE is consistent for the same individual even under a variety of conversations. These two properties lay the foundation for our idea. In the meantime, to exploit speech signals as a cue for QoE assessment is faced with a critical challenge that speech characteristics are diverse among individuals. Hence, how to construct a personalized QoE assessment model that accommodates the individual heterogeneity is one of the main focuses of our design.

3 BASIC QOE ASSESSMENT MODEL USING FEW-SHOT LEARNING

In the measurement study, we show the statistical properties of some classic features to illustrate the feasibility of our idea. The remaining question is how to extract useful features out of speech signals for QoE assessment. We employ convolutional neural network (CNN) as the classifier, with the inputs as time-frequency (t-f) domain speech spectrogram. CNN is expected to extract explicit and implicit features that are most beneficial for QoE classification from the inputs.

In the following, we first present a basic model that turns speaker's speech signals into her QoE perception via a CNN. To quickly adapt the trained model to an unseen user, few-shot learning is adopted. At the end of this section, we discuss the limitation of the basic model. To overcome it, an advanced model is developed in the next section. The basic model is vital for serving as a basis for the entire design.

3.1 CNN-based QoE Classifier

We propose to employ CNN as a classifier that maps a user's speech to her QoE perception. The input is the t-f domain speech spectrogram. Its output is a 5-point MOS, where 5 represents the best quality and 1 represents the worst.

The speech is captured by the speaker's microphone. It is then down-sampled to 16 kHz, as signals above 8 kHz barely affect the speech intelligibility and human perception [56]. A higher sampling rate may unnecessarily increase the computational complexity. The t-f domain speech spectrogram is generated by applying STFT on the time domain waveform. The STFT adopts a window size of 32 ms, hop length of 10 ms, and FFT size of 512 points under 16 kHz sampling rate, resulting in $100 \times 257 \times 1$ complex-valued scalars per second. In the implementation, we cut an audio track into multiple clips, each with a duration of 10 s. Each clip is one sample with the size $997 \times 257 \times 1$. The classifier takes the t-f domain spectrogram as input. Then, implicit features are obtained by conducting multiple levels of non-linear operations. Each operation transforms the data representation learnt at the previous level into a representation at a higher and more abstract level. In particular, the multi-layer non-linear operations, in the form of non-linear activation function and pooling layers, make the obtained data representation sensitive to subtle details embedded in the spectrogram, and insensitive to large-scale irrelevant variations resulting from speaking contextual diversity. Lastly, the learned representation is fed into the fully connected layers for QoE inference.

3.2 Handling Unseen Users via Few-shot Learning

As presented in the previous section, human speech patterns are diverse even under the same QoE perception. Hence, the CNN-based classifier should be user-specific and readily scale to an unseen user. We propose to employ the *few-shot learning* [27, 28]. Audio samples from existing users and unseen users are treated as the *source domain* and the *target domain*, respectively. The few-shot learning technique aims to fine-tune parameters of models, trained in the source domain, to adapt to the target domain within limited data samples from the target domain. To be specific, we adopt the model-agnostic meta-learning (MAML) [29], a few-shot learning framework that uses gradient descent and requires only a few gradient steps to update the model. It consists of the *meta-training* and *adaption* phases. In the following, we cover essential steps in each

phase. Although MAML is not the technical contribution of this work, it provides the structural framework of SpeechQoE.

Algorithm 1: Meta-training

input: Source dataset $\mathcal{D}_S$, learning rate hyperparameters α and β

1 $\theta_0 \leftarrow$ random initialization;
2 **while** *not finished* **do**
3 Generate a batch of tasks $\mathcal{T}$ from source dataset $\mathcal{D}_S$;
4 **for** *all generated* $\mathcal{T}_i \in \mathcal{T}$ **do**
5 $S_{\mathcal{T}_i} \leftarrow$ Sample $K \cdot M$ support instances from $\mathcal{T}_i$;
6 $Q_{\mathcal{T}_i} \leftarrow$ Sample $K \cdot M$ query instances from $\mathcal{T}_i$;
7 Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta, \mathcal{S}_{\mathcal{T}_i})$ with $S_{\mathcal{T}_i}$ via equation (2);
8 Compute $\mathcal{T}_i$-specific parameters $\theta'_{\mathcal{T}_i}$ through SGD: $\theta'_{\mathcal{T}_i} = \theta_0 - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta, \mathcal{S}_{\mathcal{T}_i})$;
9 Evaluate $\mathcal{L}_{\mathcal{T}_i}(f_{\theta'_{\mathcal{T}_i}}, Q_{\mathcal{T}_i})$ with $Q_{\mathcal{T}_i}$;
10 $\theta' \leftarrow \theta' - \beta \nabla_\theta \sum_{\mathcal{T}_i \in \mathcal{T}} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_{\mathcal{T}_i}}, Q_{\mathcal{T}_i})$;

output: Trained parameter θ'

Meta-training phase. Formally, we denote the QoE classifier as $f_\theta(\boldsymbol{x})$ with neural network parameters θ initialized to θ_0. Let $\mathcal{D}_S = \{(\boldsymbol{x}_j, y_j)\}$ be the source dataset, where $\boldsymbol{x}_j$ is the t-f spectrogram of the jth sample with y_j as its label, i.e., a discrete QoE level from 1 to 5. With $\mathcal{D}_S$, a set of tasks $\mathcal{T}$ are generated (line 3). Each task $\mathcal{T}_i \in \mathcal{T}$ is a K-shot M-way classification problem, where the classifier aims to predict M QoE classes by using K labeled instances in each class. Note that M is 5 in our case (i.e., 5 different QoE levels) and K is a small number, e.g., 5 or 10. Each task $\mathcal{T}_i$ is associated with a *support set* $S_{\mathcal{T}_i}$ and a *query set* $Q_{\mathcal{T}_i}$. The two sets are disjoint with each other ($S_{\mathcal{T}_i} \cap Q_{\mathcal{T}_i} = \emptyset$). Each set contains $K \cdot M$ instances from $\mathcal{T}_i$ (line 5-6). Each task mimics the situation that only limited labeled audio samples are available from a "virtual user" in our case. f is trained using the support set $S_{\mathcal{T}_i}$. To be specific, it computes temporary parameters $\theta'_{\mathcal{T}_i}$ via gradient descent with $S_{\mathcal{T}_i}$ (line 7-8)

$$\theta'_{\mathcal{T}_i} = \theta_0 - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta, S_{\mathcal{T}_i}), \tag{1}$$

where α is the learning rate. The loss function is defined as

$$\mathcal{L}_{\mathcal{T}_i}(f_\theta, \mathcal{S}_{\mathcal{T}_i}) = \sum_{(\boldsymbol{x_j}, y_j) \in \mathcal{S}_{\mathcal{T}_i}} y_j \log f_\theta(\boldsymbol{x_j}) + (1 - y_j) \log f_\theta(1 - \boldsymbol{x_j}). \tag{2}$$

which is a task-specific cross-entropy loss of f_θ on the support set $S_{\mathcal{T}_i}$. With the task-specific parameters $\theta'_{\mathcal{T}_i}$ for all $\mathcal{T}_i$, we then define an optimization problem, with an objective function to find the across-tasks parameters θ' that minimizes the sum of task-specific losses for all tasks in $\mathcal{T}$, $\min_\theta \sum_{\mathcal{T}_i \in \mathcal{T}} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_{\mathcal{T}_i}}, Q_{\mathcal{T}_i})$. Note that each task-specific loss $\mathcal{L}_{\mathcal{T}_i}(f_{\theta'_{\mathcal{T}_i}}, Q_{\mathcal{T}_i})$ is evaluated by the task-specific parameters $\theta'_{\mathcal{T}_i}$ on the corresponding query set $Q_{\mathcal{T}_i}$ (line 9). The optimization problem is solved by stochastic gradient descent (SGD) [45] (line 10).

$$\theta' \leftarrow \theta' - \beta \nabla_\theta \sum_{\mathcal{T}_i \in \mathcal{T}} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_{\mathcal{T}_i}}, Q_{\mathcal{T}_i})$$

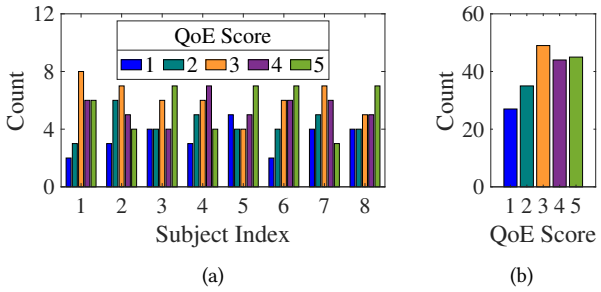


Figure 6: QoE scores are unevenly distributed. (a) Distribution for individuals. (b) Overall distribution.

where β is the learning rate for SGD optimization. The trained optimal parameters θ' then serves as an instantiation of the model f. The process of meta-training is given in Algorithm 1.

During the meta-training, each task $\mathcal{T}_i$ is generated in a way to simulate the situation that an unseen user of different speaking characteristics is encountered. The training is done on a task-basis. Hence, the model learns how to adapt to a new task (i.e., unseen user) quickly in $K \cdot M$ samples.

Adaptation. Once the model $f_{\theta'}$ is trained by the source dataset $\mathcal{D}_S$, it can be deployed and adapted to any target user. MAML ensures that the adaption can be done with only $K \cdot M$ data samples. Specifically, a new K-shot M-way classification problem is formulated with the target dataset $\mathcal{D}_T$. The original model $f_{\theta'}$ is fine-tuned to the new classification problem using a few gradient steps as

$$\tilde{\theta} \leftarrow \tilde{\theta} - \alpha \nabla_\theta \mathcal{L}_{\mathcal{D}_T}(f_{\tilde{\theta}}).$$

Note that $\tilde{\theta}$ is initialized with θ'. An annotated sample in the target dataset $\mathcal{D}_T$ is collected by asking the new user to rate her perceived quality after each voice call. Apparently, a small number of $K \cdot M$ is desirable to minimize the data collection overhead.

3.3 Limitation of the Basic Model

Although the basic model is designed to be user-specific and adapt to an unseen user quickly, it still acquires K samples from each of M classes from the unseen user. For example, when $K = 3$ (i.e., 3 labeled conversational audio tracks in each QoE level), then the model adaption under MAML demands 15 samples from the unseen user, given the 5 QoE levels. While the number is small, the sample collection overhead, in terms of time duration, may be non-negligible in practical implementation.

Figure 6a shows the QoE score distribution from 8 randomly selected subjects. 25 tests are performed for each one of them. We vary the network condition smoothly from the worst to the best. Subjects are asked to rate their perceived service quality after each test. The subjective ratings are observed highly unbalanced across five QoE levels for most individuals. Particularly, lower scores, say 1 and 2, are less reported compared with medium and higher scores. Take the second subject as an example. It takes 25 trials to gather at least 3 samples of each QoE score, which is 10 samples more than the ideal case ($25 - 3 \times 5 = 10$). This is partially because most users tend to avoid badmouthing their received service, a phenomenon that has been studied in psychological and cognitive

science [32, 66, 74]. Note that the extra 10 samples are useless for the model adaption. This situation becomes even worse if applying a larger K in few-shot learning, as more redundant samples would be collected and wasted.

One may argue to employ one-shot learning, a special case for few-shot learning, to mitigate the data collection overhead in the adaption phase. First of all, as a model tends to overfit on limited samples [71], the adoption of one-shot learning is restricted by several conditions. For example, the collected one-shot samples must satisfy the distribution of the source domain dataset [79], which cannot be guaranteed here due to user diversity. Second, the network conditions nowadays rarely down-grade to "unacceptably bad" that leads to the lowest QoE, especially for voice services that demand relatively small bandwidth. Therefore, even one-shot learning may still take a long time before acquiring a sample labeled with QoE=1[2]. In summary, it is impractical to directly apply few-shot learning to our case.

4 ADVANCED QOE ASSESSMENT MODEL

In this section, we present our proposed advanced model. It is still built on the framework of MAML-based few-shot learning, but is enhanced to quickly adapt to an unseen user within few samples which are likely to unevenly distribute across the QoE classes.

4.1 Adaption with Limited Target Samples of Uneven Distribution

Overview. The meta training phase in the advanced model is identical to that in the basic model. In the adaption phase, instead of gathering K samples in each QoE class, which may take significant time in practice, we propose to generate synthetic data for the target user. To be specific, we seek to calibrate the distribution of the target user's dataset via her limited samples. Then an adequate number of samples, covering all the QoE scores, can be sampled from the calibrated distribution. We then fine-tune the model via the joint set of real and synthetic data samples. In this way, the data collection overhead would be reduced greatly. Although the idea seems straightforward, the design is faced with two challenges. First, as only limited samples are available, they tend to be biased. It is non-trivial to infer the target user's ground truth distribution from the biased samples. Second, because of the unbalanced distribution property, it is not rare that samples of certain class(es) are missing. Without any reference, is it feasible to generate synthetic dataset for those class(es)?

To tackle these two challenges, we develop a novel data synthesizing scheme. In a holistic view, we first calibrate the distribution of the few samples of the target user by transferring statistics from the "close" source users. Then an adequate number of samples for each class of the target user are produced from the calibrated distribution. The proposed scheme is composed of three steps: *source user profiling*, *identifying close neighbors*, and *generating target data samples*.

Source user profiling. Consider a source dataset $\mathcal{D}_S$ collected from a set of source users S. Let $\mathcal{D}_p \in \mathcal{D}_S$ be the subset associated with user $p \in S$. We assume the t-f domain speech spectrogram from a user p satisfy a multivariate Gaussian distribution. Their mean vector $\boldsymbol{\mu}_p$ is expressed as $\boldsymbol{\mu}_p = \{\boldsymbol{\mu}_{p,i} | i = 1, 2, \cdots, 5\}$, where $\boldsymbol{\mu}_{p,i}$ stands for the mean vector of user's samples annotated with QoE score i. It is calculated as the mean of every single dimension in the sample vector

$$\boldsymbol{\mu}_{p,i} = \frac{1}{n_{p,i}} \sum_{j=1}^{n_{p,i}} \boldsymbol{x}_j \tag{3}$$

where $n_{p,i}$ is the number of samples of class i from user p. $\boldsymbol{x}_j$ is the t-f domain speech spectrogram of the j-th sample. Similarly, the covariance matrix is expressed as $\Sigma_p = \{\Sigma_{p,i} | i = 1, 2, \cdots, 5\}$, where $\Sigma_{p,i}$ is calculated as

$$\Sigma_{p,i} = \frac{1}{n_{p,i} - 1} \sum_{j=1}^{n_{p,i}} (\boldsymbol{x}_j - \boldsymbol{\mu}_{p,i})(\boldsymbol{x}_j - \boldsymbol{\mu}_{p,i})^\top. \tag{4}$$

Given above, the distribution of user p's t-f domain speech spectrogram can be expressed as $\mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p)$, called the *profile*.

Identifying close neighbors from source users. Once the statistics of all source users are derived, we further identify from them the ones that share the most similar profiles of the target user u. Specifically, we calculate the Euclidean distance of the sample space between the target user u and each source user $p \in S$

$$d_{u,p} = \frac{1}{n_u} \sum_{i=1}^{5} \sum_{j=1}^{n_{u,i}} ||\boldsymbol{x}_j - \boldsymbol{\mu}_{p,i}||^2. \tag{5}$$

Let $\mathcal{D}_T$ be the real samples from the target user u. n_u stands for the total number and $n_{u,i}$ is the number from each class i. Apparently, we have $n_u = \sum_i n_{u,i}$. We assume source users have samples of all QoE scores. This is a practical assumption as those samples are collected offline. The calculation of (5) does not require the presence of target samples from each class. In another word, $d_{u,p}$ can still be derived, even samples of certain QoE score(s) are absent from the target dataset. In an extreme case, as few as one sample is sufficient for the calculation. This property is desirable for our situation–the adaption no longer has to wait until the arrival of K samples from each QoE score. A total number of n_u (e.g., 3 to 5) target samples are sufficient, regardless of their classes. Apparently, it effectively shortens the period of time for adaption, comparing

Algorithm 2: Data synthesis scheme

input: Source dataset $\mathcal{D}_S$, real target dataset $\mathcal{D}_T$, source user set S

for $p \in S$ **do**
- Calculate mean vector $\boldsymbol{\mu}_{p,i}$ and $\Sigma_{p,i}$ following (3);
- Calculate covariance matrix $\Sigma_{p,i}$ following (4);
- Calculate Euclidean distance between u and p following (5);

Identify u's close neighbors from S in respect of their Euclidean distances;

Calibrate target user's profile using (6);

Generate synthetic target dataset $\tilde{\mathcal{D}}_T$ using (7);

output: Augmented target dataset $\mathcal{D}'_T = \tilde{\mathcal{D}}_T \cup \mathcal{D}_T$

[2]As the source dataset is gathered in a lab environment, data points annotated with low QoE scores can be obtained easily by creating poor network conditions via parameter setting. Hence, the above concern does not exist in data collection for meta-training.

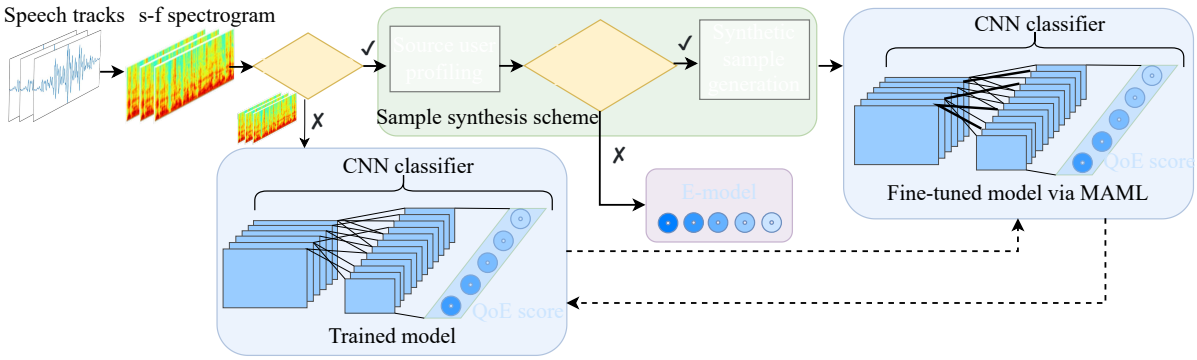


Figure 7: System architecture of SpeechQoE.

with the conventional few-shot learning. On the other hand, the value of n_u does impact the assessment accuracy, which will be evaluated in Section 5.

We adopt Euclidean distance to measure the similarity between the target user and the source user. The idea is inspired by k-nearest neighbor (k-NN) [20], a classic machine learning algorithm mainly for classification. k-NN works by finding the distances between a query sample and all the examples in the dataset, selecting the specified number examples (k) closest to the query, then votes for the most frequent label. Particularly, the Euclidean distance is commonly adopted by k-NN to quantify the similarity between two samples.

Afterwards, we select from S a subset of source users whose distances with u are less than a threshold ρ. They are considered as u's close neighbors who share high similarities in their profiles with u. Denote by S' the set of qualified neighbors. The selection of ρ and how it impacts the QoE assessment accuracy is discussed in Section 5.3.

Generating samples for the target user. We then leverage statistics from eligible neighbors to estimate the target user's data distribution. Specifically, the mean and the covariance of the distribution is calibrated as

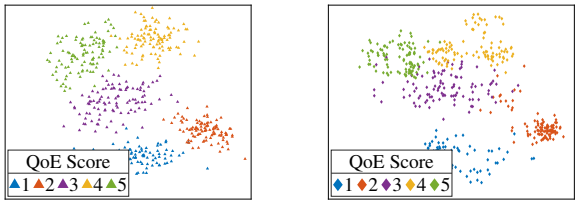
$$\boldsymbol{\mu}'_{u,i} = \frac{1}{|S'|}\sum_{p\in S'}\boldsymbol{\mu}_{p,i}, \quad \Sigma'_{u,i} = \frac{1}{|S'|}\sum_{p\in S'}\Sigma_{p,i}. \tag{6}$$

With the calibrated distribution of the target user, we are able to generate sufficient samples covering all QoE scores from the multivariate Gaussian distribution

$$\tilde{\mathcal{D}}_T = \{(\boldsymbol{x}, y)|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}'_{u,i}, \Sigma'_{u,i}),\ \forall i \in \{1, 2, \cdots, 5\}\}. \tag{7}$$

The synthetic target dataset $\tilde{\mathcal{D}}_T$, together with real dataset $\mathcal{D}_T$, then form the augmented target dataset $\mathcal{D}'_T = \tilde{\mathcal{D}}_T \cup \mathcal{D}_T$. $\mathcal{D}'_T$ is used to fine-tune the model following operations covered in Section 3.2. Theoretically, target samples can be generated as many as desired. On the other hand, the performance becomes stable as the number surpasses a certain threshold. We are going to examine it in Section 5.3. The steps of proposed sample synthesis method is given in Algorithm 2.

We perform a series of distribution tests, which determine whether our collected sample data are drawn from a certain probability distribution. In particular, the classic Kolmogorov–Smirnov (K-S) test



(a) (b)

Figure 8: t-SNE visualization of (a) synthetic samples and (b) real samples.

is adopted. A set of common statistical distributions, including Gaussian distribution, Gamma distribution, and Exponential distribution, are considered. The result suggests that our dataset, in most cases, complies with the Gaussian distribution the best among all distributions considered.

Discussion. We compare in Figure 8 the t-SNE [75] representation of synthetic samples and real samples collected from the target user. t-SNE is a statistical method to visualize high-dimensional data by giving each datapoint a location in a two or three-dimensional map. Their distributions across five QoE scores are similar, e.g., the relative positions of five clusters on the 2-D dimension. Hence, model fine-tuning through synthetic samples is expected to adapt to the target user properly. Besides, the proposed data synthesis scheme can potentially be applied to another context with a similar data hungry issue.

There have been some existing efforts on data synthesis [50, 69] and data augmentation [53, 63, 77]. Unfortunately, they are inapplicable here. Most data synthesis methods are built with specialized modules, such as Generative Adversarial Networks (GANs) [33] and Variational Autoencoders (VAEs) [82]. These methods require the design of a complex model, while our distribution calibration algorithm is simple. Besides, as they are mostly data-driven models, they require proper training before deployment. In our scenario, these models tend to be over-fitting given the limited target samples, especially with uneven distribution. Data augmentation is a conventional way of increasing the number of training samples. The mainstream approaches involve using simple tricks (e.g., cropping,

rotating, and zooming). However, they are unsuitable for t-f domain speech spectrogram, a more complex-form of data representation. More importantly, none of the above methods can deal with the absence of samples from certain class(es), a unique challenge in our problem.

4.2 Integration with Parametric Model

Our sample synthesis method relies on an adequate number of close neighbors in the source user set, so that the target user's data distribution can be accurately calibrated. In practice, this assumption may not always hold due to the lack of a large user pool in the source dataset. To address this issue, we propose to integrate with the conventional parametric model, e.g., the well-known E-model [30]. Essentially, the parametric model serves as a protective backup. It is activated once close neighbors cannot be identified in the source domain with a meaningful amount. As the adaption cannot be successfully performed, the parametric model is then employed to assess the target user's perceived QoE. Recall that a parametric model quantifies QoE through underlying network conditions. It is thus free from the restriction mentioned above. The hybrid design can effectively avoid catastrophic performance degradation caused by adaption failure. Note that the parametric model branch is not expected to execute often. Still, it helps to ensure robustness in QoE assessment. A similar idea has also been adopted in prior works [7, 81]. They combine the classic and learning-based approaches on adaptive network resource management and congestion control. Their results show that the hybrid design can significantly alleviate the issues (e.g., unseen scenarios) of clean-slate learning-based counterparts.

5 EVALUATION

5.1 Experiment Design

System setup. A VoIP testbed is set up for the experiments. Two laptops are used as terminals and installed with Linphone [3], an open-source app offering free audio/video calls with flexible configuration capabilities. To establish a connection between the two terminals, we deploy a voice service server on a desktop running Ubuntu 20.04. Asterisk [2], an open-source communication platform, is installed on the server to control the VoIP packet flow over SIP/RTP. Asterisk is the heart of this platform since every activity is through Asterisk. We configure via Linphone the destination address on each terminal as the server's IP address to get them connected. A router is deployed to facilitate the connection among the three entities. Our testbed is built on a local area network (LAN), which has no interconnection with the external Internet. The LAN environment allows the network to run as configured. As a VoIP packet will not be routed externally, it does not experience any extra congestion or delay caused by the external network. To simulate various calling environments, we control network parameters through Network Link Conditioner (NLC) [4] that is installed on both terminals. Table 1 lists the set of simulated calling environment profiles by tuning different network parameters. Specifically, packet loss rate (PLR) is the ratio of packets not received to the total number of sent packets. It is a metric representing how reliable a network is. Latency in our experiment is the one-way delay that takes to transmit a packet from one terminal to the other. We also

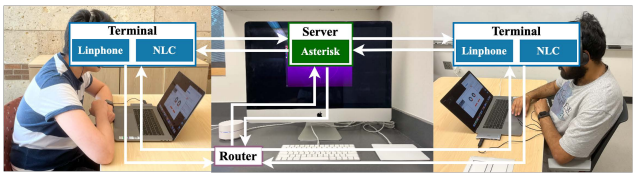


Figure 9: Testbed setup.

consider background noise, which is pre-recorded from the cafeteria. It is played during the entire calling session. Period refers to a short duration of network disconnection. It is commonly observed when there is bandwidth competition. In total, twenty-five testing conditions are created following settings in prior works [15, 19]. They are expected to cover a variety set of calling environments in real scenarios.

Table 1: Calling environment profiles.

Condition	PLR (%)	Latency (ms)	Noise (dB)	Period (s)
Excellent	[0-0.1)	[0-50)	[0-30)	[0-0.5)
Very Good	[0.1-0.5)	[50-100)	[30-50)	[0.5-1)
Good	[0.5-1.0)	[100-200)	[50-60)	[1-2)
Fair	[1.0-2.0)	[200-300)	[60-70)	[2-3)
Poor	[2.0-3.0]	[300-500]	[70-80]	[3-5]

Data collection. A total of 38 subjects, 23 male and 15 female, are recruited for the experiments. Two subjects form a pair to complete 200 calling sessions, each lasting 90 seconds. They sit in separate rooms to avoid mutual-interference. Their conversational topic is the so-called Richard's task [59], commonly employed in telephone conversational quality evaluation. It is like charades. Two subjects take turns describing a shape on a given sheet for the other to identify. During the 90 seconds conversation, they aim to guess as many shapes correctly as they can for a larger amount of rewards. Their conversation is recorded by Quick Time Player [6]. At the end of each session, subjects are asked to rate their perceived QoE from 1 to 5. To prevent subjects from excessive fatigue, the 200 sessions are accomplished in five rounds on different days. Each subject devotes about 6 hours on average to carry out the experiments. The entire data collection campaign lasts for 6 months. To our knowledge, our dataset is the first medium-scale QoE-labeled dataset for conversational voice services.

CNN classifier implementation. SpeechQoE is built with CNN. The classifier consists of four convolution layers, followed by four max pooling layers. Two fully-connected layers are attached at the end. ReLU is used as the activation function after each of the convolutional layers. L_2-regulation is employed to prevent overfitting. The classifier is trained with Adam optimizer [46]. We implement SpeechQoE using PyTorch framework [5]. The meta-training of the model is performed in a server equipped with eight NVIDIA RTX A6000 GPUs with Intel Xeon Gold-5218R 2.10 GHz processors.

We train SpeechQoE in a leave-one-out manner, in which the data collected from one subject is used as the target dataset and the data collected from all the other subjects act as the source dataset. The model training and adaption are performed under the

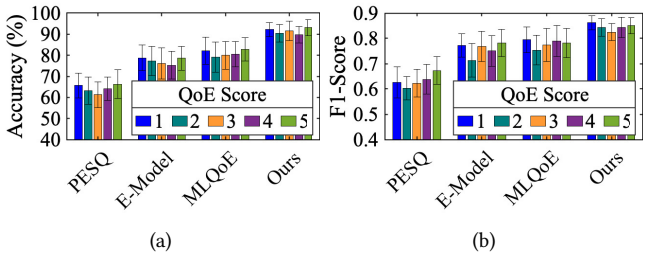


Figure 10: Performance comparison with two baseline approaches in (a) accuracy and (b) F1-score.

framework of MAML. In the training phase, we sample 74 tasks, twice of the source domain size, from the source dataset. The task is further decomposed into support set and query set, which are fed into the model for training. The adaption is performed with the target user's synthetic dataset. The inference is performed over the target user's real dataset. There are 38 evaluations in total. The overall performance is the average result from 38 evaluations.

Baseline methods. We compare SpeechQoE against the following models which represent the state-of-the-art in QoE assessment for voice services: (1) PESQ [60], a speech quality based approach. It requires the original speech source and the degraded speech files as inputs to analyze audible distortion and further convert it into a QoE score. (2) MLQoE [16], which maps network metrics, such as latency, jitter, and packet loss, into a QoE score. It employs a set of AI techniques, including Support Vector Regression (SVR), Artificial Neural Networks (ANNs), Decision Trees (DTs), and Gaussian Naive Bayes (GNB), and selects the best one automatically [73]. (3) E-model [30], which is a classical parametric model. Unlike SpeechQoE and MLQoE, which are data-driven models, E-model is an analytic model.

5.2 System-level Evaluation

Overall performance. We first compare SpeechQoE with the three baseline approaches in overall performance. Two metrics are examined, accuracy and F-1 score. Accuracy is defined as the percentage of correct assessments over all trials. The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. We observe in Figure 10 that SpeechQoE consistently outperforms the baselines for all QoE scores. Specifically, the average accuracy of PESQ, E-model, MLQoE and SpeechQoE is 64.1%, 77.2%, 80.9%, and 91.4%, respectively. The inferior performance of the baseline approaches is primarily due to the lack of subjective factors considered in their models. Instead, SpeechQoE exploits speaker's speech patterns as an essential cue to assess the subjective QoE perception. Moreover, SpeechQoE integrates E-model as a backup classifier when the speech-based classifier becomes infeasible. It thus takes advantage of both data-driven and analytic approaches in QoE modeling. Below we provide a more in-depth analysis of the results.

The result also suggests that SpeechQoE works for voice services of various qualities, from low to high. Figure 10 shows that SpeechQoE delivers consistent performance across QoE scores. Take accuracy as an example, our achieved values are 92.2% (QoE=1), 90.4% (QoE=2), 91.6% (QoE=3), 89.7% (QoE=4), 93.1% (QoE=5), respectively.

Performance breakdown. This part intends to provide in-depth understanding why SpeechQoE outperforms the baselines.

(i) User diversity. Figure 11a shows the assessment accuracy for each subject. We find that the performance variation among subjects under SpeechQoE is smaller. Specifically, the standard deviation is 3.36, 5.78, and 5.82 for SpeechQoE, E-model, and MLQoE, respectively. It indicates that our model delivers a more stable and consistent performance in QoE assessment across users than the other two. Besides, employing few-shot learning framework renders our model adaptive to user diversity. In contrast, the baselines solely rely on network conditions for QoE prediction. While MLQoE claims as a user-centric QoE model, still no subjective factors are considered. Hence, all users share a uniform model, which can hardly capture user's uniqueness in their perception.

Table 2: QoE assessment on two participant groups.

QoE Score	1	2	3	4	5
Group 1	0.92	0.91	0.90	0.89	0.92
Group 2	0.92	0.90	0.92	0.90	0.93

We further evaluate SpeechQoE on two participant groups. One is non-college students (group 1), whereas the other is college students (group 2). We cannot identify noticeable difference from the results of the two groups, with their average accuracy as 91.4% and 90.9%, respectively. It indicates that our scheme works for users of diverse background.

(ii) Background noise. In practice, calls are often made in noisy environments, e.g., cafes, streets, and shopping malls. A QoE model should account for this influence factor as well. In the experiments, we play pre-recorded sound in testing rooms to create a noisy environment. Three sound levels are adopted, i.e., low (0-50 dB), medium (50-60 dB), and high (60-80 dB). Figure 11b shows that the performance of baseline approaches degrades significantly with a higher background noise, which apparently causes non-negligible influence on speaker's QoE perception during a call. Note that this factor has been largely overlooked in most parametric models. On the other hand, noisy environments would impact the way people speak. For example, they may unconsciously slow down their speech rate and raise their voices. Those minute changes can be effectively captured by our SpeechQoE.

(iii) Quality perception of the peer. In this series of experiments, we set different network conditions at the two calling parties. Three settings are considered. In case 1, calling environments are set as *excellent* (in in Table 1) for both terminals. In case 2, one of them is set as *excellent*, whereas the other is set as *good*. In case 3, they are set as *excellent* and *poor*, separately. We aim to simulate "symmetric" and "asymmetric" uplink/downlink conditions via these settings. We observe that SpeechQoE performs consistently over the three cases. However, the accuracy of baselines drops significantly as network conditions become asymmetric. In a conversational call, one party's quality perception is also dependent on that of the peer. Imagine that the callee keeps on asking the caller to repeat due to the poor connection at the callee side, while the caller can hear the callee clearly. Neither of them would rate the QoE high at the end

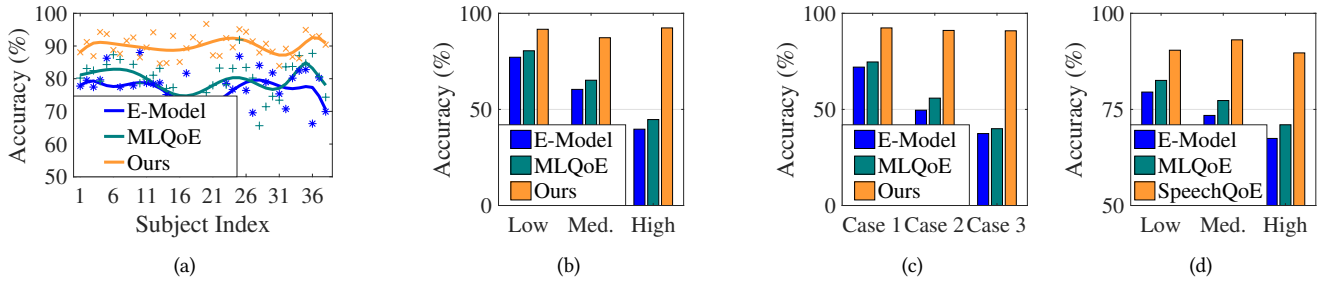


Figure 11: Performance comparison under various impact factors. (a) User diversity, (b) background noise, (c) quality perception of the peer, and (d) call fatigue.

of the call. This situation can be counted by asymmetric network conditions experienced by the uplink and downlink. As parametric models treat calling parties separately, it would wrongly classify caller's QoE and thus impair the overall accuracy performance. On the contrary, our approach predicts QoE through speech patterns, which nicely reflect the speaker's in-situ perception, including the influence introduced by the peer.

(iv) Call fatigue. Call fatigue can also impact QoE. A similar phenomenon has been observed in other online applications and services, such as web browsing and watching videos [57, 67]. It is thus crucial to count this factor too. In the experiment, participants are asked to evaluate their fatigue level, from low to high, after each calling session. We notice that a user's standard for QoE rating is dynamic, subject to her fatigue status. Hence, it is crucial for a QoE model to capture such dynamics. Figure 11d shows that the performance of our scheme is relatively stable across all conditions. It implies that speech serves as a nice QoE indicator robust to speaker's fatigue levels. On the other hand, the performance of baselines degrades with respect to the increased fatigue. As discussed, neither E-model or MLQoE considers subjective factors in their modeling. As a result, the derived models are static regardless of user's spiritual conditions.

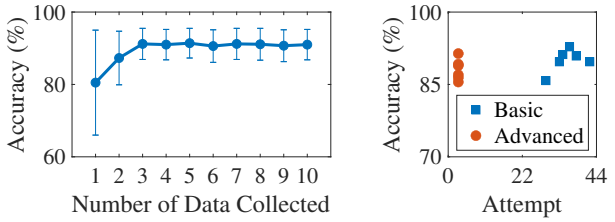


Figure 12: QoE assessment accuracy vs. data collection overhead. (a) The number if real samples needed by SpeechQoE for model adaption. (b) Comparison between the advanced model and basic model.

Adaption overhead. This part is to evaluate the adaption phase of the advanced scheme.

(i) Number of samples needed. As one of the contributions of this work, we develop a data synthesis scheme so that the number of real samples collected from the target user can be reduced. Figure 12a shows the assessment accuracy by varying the sample number

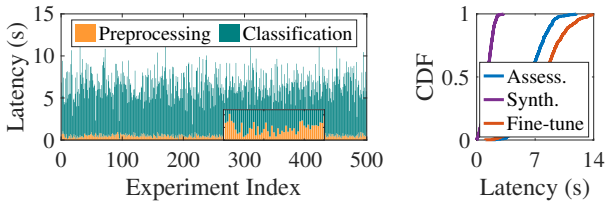


Figure 13: QoE assessment latency. (a) Stacked latency for QoE assessment when adaption is unnecessary. (b) CDF.

from 1 to 10. The accuracy is enhanced by a large margin as the number increases from 1 to 3. After that, the performance becomes stable. It implies that our scheme calibrates the target user's profile accurately with as few as 3 real samples. The result is promising because the original model can be adapted to a new user with rather minor data collection overhead.

To evaluate the effectiveness of data synthesis, we also implement the basic model, which employs the conventional K-shot learning. Specifically, Figure 12b is a performance comparison between the basic scheme (i.e., without data synthesis) and the advanced scheme (i.e., with data synthesis). Specifically, the basic model needs a total of 30-45 samples to achieve the same accuracy as the advanced model when $K = 5$. This is mainly attributed to the uneven distribution of sample classes as discussed in Section 3.3; lower ratings are less seen than the others. In order to gather at least 5 samples for each class, more calling sessions are needed, as many of them would produce redundant samples. As for SpeechQoE, as few as 3 samples, 10 times less than the basic model, are needed from the new user to deliver the optimum performance. The result suggests that the advanced model is more practical for deployment than the basic model (with the conventional few-shot learning as the substrate), due to its negligible data collection overhead.

(ii) QoE assessment latency. We now examine the time it takes to assess QoE. Unlike model training, which is done on the GPU server, the assessment is performed on personal terminals (i.e., laptops with Intel Core i7 2.3GHz processors). We first consider the case when adaption is unnecessary, i.e., the assessment is done directly using the trained model. The latency is mainly attributed to audio preprocessing and inference. Figure 13a exhibits the latency over 500 trials. The total value is 6.51 s on average, including 0.5 s on

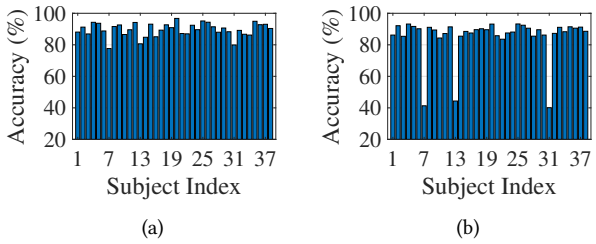


Figure 14: Assessment accuracy for individuals (a) with E-model integrated and (b) without E-model integrated.

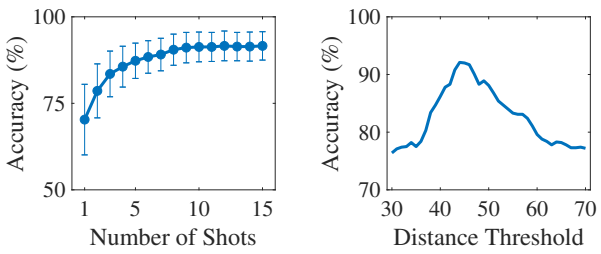


Figure 15: Impact of the no. of shots in the adaption.

Figure 16: Impact of distance threshold ρ.

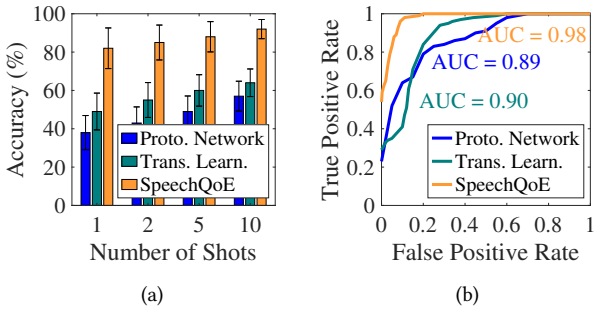


Figure 17: Performance comparison with the state-of-the-art learning frameworks. (a) Accuracy (b) RoC curve.

preprocessing and 6.11 s on testing. Figure 13b further illustrates the CDF; 90% of measurements are lower than 8.77 s. We then consider the case when adaption is needed. Since the preprocessing and inference operations are identical to the first case above, we focus on evaluating the time needed for the adaption. It is triggered once a new user arrives. It involves two steps: data synthetization and model fine-tuning. Figure 13b shows that the latency for data synthetization and model fine-tuning is 1.29 s and 7.84 s on average, respectively. The total latency for adaption is 9.13 s on average. Note that model adaption is needed only once for a new user. In summary, a predicted QoE score will be available shortly after a conversational call in most cases.

In the current setting, SpeechQoE produces one QoE score after a call finishes. This score reflects the speaker's perceived quality during the entire call. It is practically acceptable for service providers to have the QoE inference result ready in several seconds. Take the current audio/video telecommunication apps as an example. After each call, users are typically prompted with a post-service survey to rate the service quality, for example, from one star to five stars corresponding to a QoE score from 1 to 5. This manual process would not be faster than ours. Besides, many users may be reluctant to provide their ratings as such process is effort-demanding and sometimes annoying. As for SpeechQoE, the assessment is automatically carried out. It avoids bothering users with questions to collect opinions and feedback, and significantly reduces human labor efforts.

5.3 SpeechQoE Deep Dive

Integration of parametric model. We show in Figure 14a and 14b the QoE prediction accuracy with and without integrating a parametric model into our design, respectively. E-model is applied in the implementation. We notice in Figure 14a that the accuracy for each individual is relatively consistent. However, the accuracy degrades drastically for three subjects (with indices 7, 13, and 31) in Figure 14b. This is because insufficient close neighbors are identified in the source user set (i.e., the remaining 37 subjects). Consequently, the model fine-tuning cannot be performed successfully. Although the accuracy produced by the E-model for the three subjects is around 80%, it effectively avoids catastrophic situations, an issue widely observed in data-driven approaches. Note that our design is modular. It is convenient to replace the E-model with any other parametric or speech quality based model in the implementation. We are going to evaluate those combinations in the future.

Impact of number of shots. SpeechQoE adopts the framework of MAML. It is important to evaluate the impact of the number of shots (K) on the assessment accuracy. Figure 15 gives the result by varying K. The accuracy increases as K grows from 1 to 8 and stays slightly above 90% after that. Therefore, we can safely adopt 8-shot learning to achieve a satisfactory assessment accuracy. Note that we do not need to collect all 40 samples ($8 \times 5 = 40$) from the new user for model adaption. As implied by Figure 12a, as few as 3 real samples are sufficient to calibrate the new user's *profile*. The remaining 37 samples are all synthetic from the calibrated distribution. It thus largely alleviates the data collection overhead.

Threshold ρ for close neighbor selection. In the proposed data synthesis scheme, the threshold ρ plays an important role. A target user's close neighbors are the source users with their Euclidean distances (to the target user) smaller than ρ. The target user's profile is then derived from the statistics of those neighbors. Figure 16 shows the impact of ρ on the assessment accuracy. We find that the optimum value exists at around 45 in our implementation. Either a too low or a too high ρ leads to inferior performance. Specifically, a smaller ρ implies a more stringent neighbor selection rule. On one hand, speech patterns from the selected neighbors are more similar to that of the target user. On the other hand, fewer neighbors are qualified. In essence, we need to strike a balance between these two factors.

Performance comparison with other learning frameworks. SpeechQoE adopts MAML to adapt the model to unseen users. Here we compare two other learning frameworks, transfer learning and prototypical network, which can be applied for a similar purpose. Regarding transfer learning, we follow the prevalent approach [79]. In the adaption, we freeze convolutional layers and retrain the FC

layer to fine-tune the model using samples from the target user. Prototypical network [71] as another representative in meta-learning [68], learns a metric space in which classification can be performed by computing distances to prototype representations of each class. Figure 17a shows the accuracy with respect to the number of shots. Overall, SpeechQoE outperforms the other two in all cases, especially when fewer shots are adopted. Hence, SpeechQoE is more suitable for our scenario where an efficient adaption is desired. Figure 17b depicts the receiver operating characteristic (ROC) curves. We also specify the area under curve (AUC) for each approach. AUC is the measure of the ability of a classifier to distinguish between classes. Typically, a higher AUC is desired. SpeechQoE delivers the best classification performance among the three.

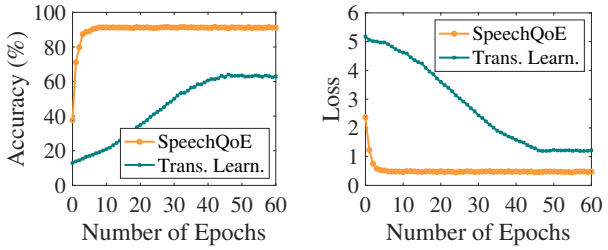


Figure 18: Training overhead in the adaption. (a) Accuracy vs. number of epochs. (b) Loss vs. number of epochs.

Training overhead in adaption. Figure 18a plots the accuracy versus the number of epochs for fine-tuning. One epoch is when all samples in the support set of target dataset are passed both forward and backward through the model once. The accuracy becomes stable after 10 epochs in our approach. Nonetheless, it takes more than 40 epochs for the transfer learning [79] to reach stability. A similar observation is obtained in Figure 18b. We conclude that few-shot learning is a more desirable learning framework than transfer learning for a smaller training overhead during model adaption.

6 RELATED WORK

Existing QoE assessment methods for voice services can be broadly classified into the following three categories.

Parametric models. These models quantify QoE through the characterization of underlying networks. A broad set of dedicated parameters are employed, such as source rate, packet loss, coding scheme, delay, and jitter. The E-model, defined by the ITU-T in Rec. G.107 [30], is a classic parametric model that provides estimation for the overall quality in a voice conversation as a weighted linear sum over 20 input parameters. Despite its popularity, E-model is static in a sense that all parameters are fixed. Therefore, it is mostly used at the network planning stage, rather than quantifying QoE at *runtime* for live services. To bridge the gap, some of its extensions have been developed to adapt to network dynamics and are thus more suitable for online assessment [12, 15]. All these methods express the relation between network parameters and QoE into mathematical formulas explicitly. More recently, some research proposes to leverage artificial intelligent techniques to explore the implicit and complex relationships [8, 14, 16, 19, 70]. Machine learning models, such as artificial neural networks [14], Bayesisan network [8], logistic regression [19], and support vector machine (SVM) [70], have been employed.

All these models primarily make use of network-centric quality of service (QoS) to infer QoE. As pointed out by many prior works (e.g., [39, 41, 42]), the former does not necessarily reflect the ground truth of the latter in many cases. Besides, as user diversity is neglected, they are incapable of providing personalized QoE assessment, the main focus of this work.

Speech quality based models. They estimate QoE by analyzing speech quality without specific knowledge of underlying network conditions. They can be further classified as full-reference models [60–62, 76] and reference-free models [26, 43, 55, 58]. In the former, both the original speech source and the degraded speech file are required. The audible distortion is then converted to a QoE score. PESQ [60], as defined by ITU-T Rec. P.862, falls into this category. The need for the original speech signals makes full-reference models impractical assessing the quality of voice services in real time. As opposed to the full-reference models, reference-free models directly extract distortion measures from degraded speech to derive the final score. Among them, ITU-T Rec. P.563 [58] serves as the state-of-the-art algorithm. This line of research focuses on reconstructing a clean reference signal from the degraded signal and then calculating the distortion. Techniques, such as noise estimation and signal recovery, are employed.

Being another objective factor based models, the above approaches bear a similar concern as the parametric model–the derived model is uniform for all users and inapplicable for personalized QoE assessment. Although also relying on speech analysis, the philosophy of our design is totally different: First, we treat speech as a physiological marker to reveal speaker's perceived quality of voice services, rather than an objective indicator of underlying network conditions. Second, instead of quantifying speech quality and remedying signal distortion, we look into the way a user speaks, e.g., speed, pitch, rhythm, loudness, tone, etc., as impacted by the service quality.

Psychophysiology-based models. Psychophysiology is concerned with the measurement of physiological signals and psychological correlates thereof. It relies on the captured physiological data along with the psychological bases of perceptual and cognitive processes. Some recent works investigate the feasibility of utilizing physiological data, such as facial expression, electrodermal activity (EDA), electrocardiography (ECG), and electroencephalogram (EEG) readings, for QoE assessment in audio-visual entertainment [10, 11, 48, 65]. For example, Porcu *et al.* [65] mapped user's facial expressions and gaze direction to perceived quality for watching videos. An SVM with a quadratic kernel and a k-NN classifier are employed. Liu *et al.* [52] explored eye-tracking data to estimate visual attention and further combined with other features extracted from images to assess the perceived quality of images. Lassalle *et al.* [49] study human perception of video quality through subjective assessment and physiological measurement, such as blood volume pulse, skin conductance, and eye tracking.

The access of physiological signals requires designated sensors that are absent from most PCs and personal terminals. Therefore, the deployment of the above models is largely confined by the physiological signal accessibility. As a result, many of them are

more meaningful for laboratory testing, rather than regular usage of commercial voice services. In contrast, our design utilizes user's speech which is readily accessible via microphones in most personal devices[3]. It is thus more practical for wide adoption.

7 DISCUSSION AND FUTURE WORK

Privacy considerations. As a user's speech contains rich information for identity recognition, it is critical to ensure that the proposed SpeechQoE does not cause any privacy disclosure to the user. In our case, an original model, trained offline by the source dataset, is pre-installed on the terminals together with the voice service app. The adaption phase is triggered by the arrival of a new user. Several conversational calls are recorded by the terminal during the adaption phase. Their corresponding t-f domain spectrograms serve as target samples for model fine-tuning. All operations involved are performed locally. As no data leaves the terminal, user's privacy is well preserved. Besides, the terminal's mic does the recording, audio tracks only include voice from the enrolled user, without the peer. It thus completely avoids any potential law violation in recording other's oral communication without permission[4]. To facilitate model adaption, a bunch of source users' profiles are pre-loaded to the terminal with the original model. Note that this information has been anomynized. More importantly, profiles are highly abstract statistics, i.e., mean and covariance of source user's samples. It is impossible to convert them back to individual audio tracks.

Impact from other factors. Other than perceived service quality, speech patterns may be affected by other factors, for example, a speaker's mood in a call. A person may unconsciously speak faster when announcing exciting news; a speech tone may be lowered when she feels depressed. Apparently, the vocal variance caused by these factors should be eliminated from the model. In general, our idea is to suppress features in speech signals that are also subject to other impact factors, while amplify the ones that are primarily affected by the service quality. One viable solution is to enhance the classifier by applying a Siamese network [17]. The idea of the Siamese network is to employ twin substructures with the same neural network and weights. The whole model is trained through sample pairs: two samples with the same QoE label but associated with different moods. Each element from a pair is passed through each of the two substructures separately. The model is trained in a way that two substructures cannot distinguish between such two samples in testing. The designed structure, together with the training process, allow the model to tolerate inconsistency in speech patterns caused by speaker's mood. We plan to incorporate this idea into the design in our future work. Note that Siamese network has been widely adopted in model training with feature reconstruction for being good at distilling relevant features and eliminating the distracting ones [18, 31, 78].

Implementing SpeechQoE on resource-restricted terminals. We have so far tested SpeechQoE on regular laptops. The main operations involved are model adaption and inference, whereas the entire meta-training happens offline. According to the results discussed in Section 5.2, they can be executed in several seconds. Realizing that conversational calls are also widely carried on mobile terminals, it is equivalently important to achieve practical performances on them too, especially the resource-restricted ones. The objective is to shrink the size of the CNN-based classifier without causing noticeable accuracy degradation. Existing approaches include using reduced precision [22, 44, 80] (e.g., round the original model parameters in 32-bit floating point to 8-bit integer) and weight pruning [36–38, 54] (e.g., prune the non-essential near-zero weights after training). Besides, a large and complex teacher network can be also used to train a small student network for comparable results, thus distilling the knowledge to run the small network on mobile devices [13, 40]. The above-mentioned methods are applied after meta-learning in our case. Hence, a shrunk model is installed on resource-restricted terminals. Given the significantly reduced size, adaption and inference are envisioned to perform much more efficiently.

Hardware heterogeneity. The auditory property of two speakers/mics can be different. Consequently, hardware heterogeneity might affect the model performance. To address this issue, one feasible solution is to apply the domain adaption to the basic model when a user switches to a new terminal. Like how the scheme handles an unseen user, the framework of MAML can be employed. To execute domain adaption, a few labeled speech samples should be collected from the new device. Then the model is fine-tuned using these samples to adapt to the new device. We plan to investigate this idea in our future work.

8 CONCLUSION

In this paper, we demonstrate that speech signals can serve as a new type of indicator to infer QoE in voice services. We develop SpeechQoE, a personalized QoE assessment model to convert speech-based cues into a QoE score. The model is built with CNN classifier. In order to adapt the model to a new user, we adopt the framework of MAML. A lightweight data synthesis scheme is developed to mitigate the data collection overhead for model adaption. SpeechQoE is a hybrid design that integrates with the E-model; it is effective to avoid drastic performance degradation caused by insufficient "close neighbors" in the user pool. Comprehensive experiments are executed to evaluate the performance of SpeechQoE. It delivers satisfactory performance under a variety of settings. Moreover, the model can quickly adapt to a new user with as few as 3 samples. In summary, SpeechQoE potentially opens up a new direction for harnessing speech sensing for personalized QoE assessment. The proposed lightweight data synthesis scheme is applicable to another context with similar data-hungry issues.

[3]There is no need of QoE assessment of voice services on devices without microphones.
[4]Under the federal Wiretap Act [1], it is illegal for any person to secretly record an oral, telephonic, or electronic communication that other parties to the communication reasonably expect to be private.

REFERENCES

[1] 1986. Electronic Communications Privacy Act of 1986. https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/1285#14rkti.
[2] 2021. Asterisk. https://www.asterisk.org.
[3] 2021. Linphone. https://www.linphone.org.
[4] 2021. Network Link Conditioner. https://developer.apple.com/download/all/.
[5] 2021. PyTorch. https://pytorch.org.
[6] 2021. Quick Time Player. https://support.apple.com/downloads/quicktime.
[7] Soheil Abbasloo, Chen-Yu Yen, and H Jonathan Chao. 2020. Classic meets modern: A pragmatic learning-based congestion control for the internet. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication (SIGCOMM)*. 632–647.

[8] Abdulraqeb Alhammadi, Ayman El-Saleh, and Ibraheem Shayea. 2021. MOS Prediction for Mobile Broadband Networks Using Bayesian Artificial Intelligence. In *2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST)*. IEEE, 47–50.

[9] Murray Alpert, Enrique R Pouget, and Raul R Silva. 2001. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders* 66, 1 (2001), 59–69.

[10] Jan-Niklas Antons, Robert Schleicher, Sebastian Arndt, Sebastian Moller, Anne K. Porbadnigk, and Gabriel Curio. 2012. Analyzing Speech Quality Perception Using Electroencephalography. *IEEE Journal of Selected Topics in Signal Processing* 6, 6 (2012), 721–731.

[11] Sebastian Arndt, Jan-Niklas Antons, Robert Schleicher, and Sebastian Möller. 2016. Using electroencephalography to analyze sleepiness due to low-quality audiovisual stimuli. *Signal Processing: Image Communication* 42 (2016), 120–129.

[12] Haytham Assem, Mohamed Adel, Brendan Jennings, David Malone, Jonathan Dunne, and Pat O'Sullivan. 2013. Online estimation of VVoIP Quality-of-Experience via network emulation. (2013).

[13] Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? *Advances in neural information processing systems* 27 (2014).

[14] Abdelhak Bentaleb, Saad Harous, et al. 2021. Video QoE Inference with Machine Learning. In *2021 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 1048–1053.

[15] Prasad Calyam, Eylem Ekici, Chang-Gun Lee, Mark Haffner, and Nathan Howes. 2007. A "GAP-model" based framework for online VVoIP QoE measurement. *Journal of communications and networks* 9, 4 (2007), 446–456.

[16] Paulos Charonyktakis, Maria Plakia, Ioannis Tsamardinos, and Maria Papadopouli. 2016. On User-Centric Modular QoE Prediction for VoIP Based on Machine-Learning Algorithms. *IEEE Transactions on Mobile Computing* 15, 6 (2016), 1443–1456.

[17] S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 539–546 vol. 1.

[18] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 539–546.

[19] Elena Cipressi and Maria Luisa Merani. 2020. An Effective Machine Learning (ML) Approach to Quality Assessment of Voice Over IP (VoIP) Calls. *IEEE Networking Letters* 2, 2 (2020), 90–94.

[20] Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory* 13, 1 (1967), 21–27.

[21] Charles Darwin. 2015. The expression of the emotions in man and animals. In *The expression of the emotions in man and animals*.

[22] Roberto DiCecco, Lin Sun, and Paul Chow. 2017. FPGA-based training of convolutional neural networks with a reduced precision floating-point library. In *2017 International Conference on Field Programmable Technology (ICFPT)*. IEEE, 239–242.

[23] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[24] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. 1459–1462.

[25] Grant Fairbanks and Wilbert Pronovost. 1938. Vocal pitch during simulated emotion. *Science* 88, 2286 (1938), 382–383.

[26] T.H. Falk and Wai-Yip Chan. 2006. Nonintrusive speech quality estimation using Gaussian mixture models. *IEEE Signal Processing Letters* 13, 2 (2006), 108–111.

[27] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28, 4 (2006), 594–611.

[28] Michael Fink. 2004. Object classification from a single example utilizing class relevance metrics. *Advances in neural information processing systems* 17 (2004).

[29] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic Meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning*.

[30] ITU-T Rec. G.107. 2005. *The E-model, a computational model for use in transmission planning*.

[31] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *Advances in neural information processing systems* 31 (2018).

[32] Thomas Gilovich, Dale Griffin, Daniel Kahneman, et al. 2002. *Heuristics and biases: The psychology of intuitive judgment*.

[33] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. In *Proceedings of the Neural Information Processing Systems*.

[34] Ashishkumar Prabhakar Gudmalwar, Ch V Rama Rao, and Anirban Dutta. 2019. Improving the performance of the speaker emotion recognition based on low dimension prosody features vector. *International Journal of Speech Technology* 22, 3 (2019), 521–531.

[35] Kurt Hammerschmidt and Uwe Jürgens. 2007. Acoustical correlates of affective prosody. *Journal of voice* 21, 5 (2007), 531–540.

[36] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[37] Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems* 28 (2015).

[38] Yihui He, Xiangyu Zhang, and Jian Sun. 2017. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*. 1389–1397.

[39] A Hesam Mohseni, Amir Hossein Jahangir, and Seyed Mohammad Hosseini. 2021. Toward a comprehensive subjective evaluation of VoIP users' quality of experience (QoE): a case study on Persian language. *Multimedia Tools and Applications* 80, 21 (2021), 31783–31802.

[40] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[41] Rahul Jaiswal. 2021. Influence of Silence and Noise Filtering on Speech Quality Monitoring. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 109–113.

[42] Rahul Kumar Jaiswal and Rajesh Kumar Dubey. 2022. CAQoE: A Novel No-Reference Context-Aware Speech Quality Prediction Metric. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).

[43] Chiyi Jin and R. Kubichek. 1996. Vector quantization techniques for output-based objective speech quality. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 1. 491–494 vol. 1.

[44] Patrick Judd, Jorge Albericio, Tayler Hetherington, Tor Aamodt, Natalie Enright Jerger, Raquel Urtasun, and Andreas Moshovos. 2015. Reduced-precision strategies for bounded memory in deep neural nets. *arXiv preprint arXiv:1511.05236* (2015).

[45] Jack Kiefer and Jacob Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics* (1952), 462–466.

[46] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[47] Ernest Kramer. 1964. Elimination of verbal cues in judgments of emotion from voice. *The Journal of Abnormal and Social Psychology* 68, 4 (1964), 390.

[48] Eleni Kroupi, Philippe Hanhart, Jong-Seok Lee, Martin Rerabek, and Touradj Ebrahimi. 2014. Predicting subjective sensation of reality during multimedia consumption based on EEG and peripheral physiological signals. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[49] Julie Lassalle, Laetitia Gros, and Gilles Coppin. 2011. Combination of physiological and subjective measures to assess quality of experience for audiovisual technologies. In *2011 Third international workshop on quality of multimedia experience*. IEEE, 13–18.

[50] Jaehoon Lee, Jihyeon Hyeong, Jinsung Jeon, Noseong Park, and Jihoon Cho. 2021. Invertible Tabular GANs: Killing Two Birds with One Stone for Tabular Data Synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 4263–4273.

[51] Herman Levin and William Lord. 1975. Speech pitch frequency as an emotional state indicator. *IEEE Transactions on Systems, Man, and Cybernetics* 2 (1975), 259–273.

[52] Hantao Liu and Ingrid Heynderickx. 2011. Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 7, 971–982.

[53] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. 2020. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2970–2979.

[54] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. 2017. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*. 2736–2744.

[55] L. Malfait, J. Berger, and M. Kastner. 2006. P.563—The ITU-T Standard for Single-Ended Speech Quality Assessment. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 6 (2006), 1924–1934.

[56] Brian B. Monson, Eric J. Hunter, Andrew J. Lotto, and Brad H. Story. 2014. The perceptual significance of high-frequency energy in the human voice. *Frontiers in Psychology* 5 (2014).

[57] Marta Orduna, Pablo Pérez, César Díaz, and Narciso García. 2020. Evaluating the influence of the HMD, usability, and fatigue in 360VR video quality assessments. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 682–683.

[58] ITU-T Rec. P.563. 2004. *Single-ended method for objective speech quality assessment in narrow-band telephony applications*.

[59] ITU-T Rec. P.805. 2007. *Subjective evaluation of conversational quality.*

[60] ITU-T Rec. P.862. 2001. *Perceptual evaluation of speech quality: an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.*

[61] ITU-T Rec. P.862.2. 2005. *Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs.*

[62] ITU-T Rec. P.863. 2011. *Perceptual objective listening quality assessment,.*

[63] Seong-Jin Park, Seungju Han, Ji-Won Baek, Insoo Kim, Juhwan Song, Hae Beom Lee, Jae-Joon Han, and Sung Ju Hwang. 2020. Meta variance transfer: Learning to augment from the others. In *International Conference on Machine Learning*. PMLR, 7510–7520.

[64] Tim Polzehl, Alexander Schmitt, Florian Metze, and Michael Wagner. 2011. Anger recognition in speech using acoustic and linguistic cues. *Speech Communication* 53, 9-10 (2011), 1198–1209.

[65] Simone Porcu, Alessandro Floris, Jan-Niklas Voigt-Antons, Luigi Atzori, and Sebastian Möller. 2020. Estimation of the quality of experience during video streaming from facial expression and gaze direction. *IEEE Transactions on Network and Service Management* 17, 4 (2020), 2702–2716.

[66] Emily Pronin. 2007. Perception and misperception of bias in human judgment. *Trends in cognitive sciences* 11, 1 (2007), 37–43.

[67] Raimund Schatz, Sebastian Egger, and Kathrin Masuch. 2012. The impact of test duration on user fatigue and reliability of subjective quality ratings. *Journal of the Audio Engineering Society* 60, 1/2 (2012), 63–73.

[68] Jürgen Schmidhuber. 1987. *Evolutionary Principles in Self-referential Learning: On Learning how to Learn: the Meta-meta-meta...-hook.* Diploma Thesis. Technische Universitat Munchen, Germany.

[69] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. *Advances in neural information processing systems* 31 (2018).

[70] Susanna Schwarzmann, Clarissa Cassales Marquezan, Riccardo Trivisonno, Shinichi Nakajima, Vincent Barriac, and Thomas Zinner. 2022. ML-based QoE Estimation in 5G Networks Using Different Regression Techniques. *IEEE Transactions on Network and Service Management* (2022).

[71] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30 (2017).

[72] William F Soskin and Paul E Kauffman. 1961. Judgment of emotion in word-free voice samples. *Journal of Communication* (1961).

[73] Ioannis Tsamardinos, Amin Rakhshani, and Vincenzo Lagani. 2015. Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. *International Journal on Artificial Intelligence Tools* 24, 05 (2015), 1540023.

[74] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185, 4157 (1974), 1124–1131.

[75] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[76] Marcel Waltermann, Alexander Raake, and S. Moller. 2010. Quality Dimensions of Narrowband and Wideband Speech Transmission. *Acta Acustica united with Acustica* 96 (11 2010), 1090–1103.

[77] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. 2018. Low-Shot learning from imaginary data. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7278–7286.

[78] Xiangyu Xu, Jiadi Yu, Yingying chen, Qin Hua, Yanmin Zhu, Yi-Chao Chen, and Minglu Li. 2020. TouchPass: Towards Behavior-Irrelevant on-Touch User Authentication on Smartphones Leveraging Vibrations. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. Article 24, 13 pages.

[79] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How Transferable Are Features in Deep Neural Networks?. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. 3320–3328.

[80] Ye Yu and Niraj K Jha. 2020. SPRING: A sparsity-aware reduced-precision monolithic 3D CNN accelerator architecture for training and inference. *IEEE Transactions on Emerging Topics in Computing* (2020).

[81] Huanhuan Zhang, Anfu Zhou, Jiamin Lu, Ruoxuan Ma, Yuhan Hu, Cong Li, Xinyu Zhang, Huadong Ma, and Xiaojiang Chen. 2020. OnRL: improving mobile video telephony via online reinforcement learning. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[82] Jian Zhang, Chenglong Zhao, Bingbing Ni, Minghao Xu, and Xiaokang Yang. 2019. Variational Few-Shot Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.