CrowdMi: Scalable and Diagnosable Mobile Voice Quality Assessment Through Wireless Analytics

Ye Ouyang, Tan Yan, and Guiling Wang

***Abstract*—Scalable and diagnosable are the two most crucial needs for voice call quality assessment in mobile networks. However, while these two requirements are widely accepted by mobile carriers, they do not receive enough attention during the development. Current related research mainly focuses on audio feature analysis, which is costly, sensitive to language and tones, and infeasible to be applied to large-scale mobile networks. In this paper, we revisit this problem, and for the first time explore *wireless network*, the causal factor that directly impacts the mobile voice quality but yet lacks attention for decades. We design CrowdMi, a wireless analytical tool that model the mobile voice quality by crowdsourcing and mining the network indicators of cellphones. CrowdMi mines hundreds of network indicators to build a causal relationship between voice quality and network conditions, and carefully calibrates the model according to the widely accepted perceptual objective listening quality assessment (POLQA) voice assessment standard. We implement a light-load CrowdMi Client App in Android smartphones, which automatically collects data through user crowdsourcing and outputs to the CrowdMi Server in our data center that runs the mining algorithm. We conduct a pilot trial in VoLTE network in different geographical areas and network coverages. The trial shows that the CrowdMi does not require any additional hardware or human effort, and has very high model accuracy and strong diagnosability.**

***Index Terms*—Crowdsourcing, data mining, LTE, voice quality.**

I. Introduction

SMARTPHONES penetrate into people's life in a gallop, committing to provide better connectivity, and more importantly, higher quality voices to people. Despite tons of new Apps invented every year, voice call remains the most important and serious activity among all the cellphone usages. In 2013, on average, people spend 39 min daily on phone calls, and prefer to use voice call to carry time-sensitive content [1]. In addition to the substantial role the phone call acts in people's social life, the voice call itself usually is operated outdoor with high noise and less delay tolerable to human's perception. All these facts make the quality of voice calls always the most important metric to evaluate the comprehensive quality of a mobile network, and the performance indicator that mobile carrier always need to assess with highest priority [2], [3].

Current research in mobile voice call quality assessment mainly focuses on directly evaluating the audio quality of the speech. Perceptual objective listening quality assessment (POLQA) [4] is a standard provided by ITU-T that takes the audio clips as the input and compares it with prerecorded reference speeches to provide objective voice quality evaluation. Under such architecture, different models are built in [5]–[11] to model audio quality using signal features. Analysis on the audio quality to human perception is conducted in [12] and [13] that extracts the key features that are highly related to human perception. To assess the quality of human voice, features in different languages and tones are considered in [14]–[16].

While audio features may directly reflect voice call quality, evaluating them is very cumbersome and costly. To ensure noiseless testing environment, each single evaluation requires professional hardwares, such as high-definition recorders, headphones, and playbacks, carefully configured and operated by domain experts [4]. Even with lossless audio, the testing results usually are subjective to languages and tones [17]. Moreover, the evaluation of audio quality does not analyze the root cause of the change of the quality, and thus is unable to provide a guideline for mobile system diagnose and optimization. On the other hand, however, mobile carriers' ultimate interest is a large-scale voice quality assessment for their network and localization of possible issues. The existing voice quality evaluation definitely cannot satisfy this interest, even for cell-level assessment. Thus, a feasible voice assessment method providing causal analysis for mobile carrier is heavily demanded. In other words, such method needs to fulfill the following two requirements: 1) *scalable*, the evaluation needs to involve as less hardware and human efforts as possible and 2) *diagnosable*, the evaluation results must be interpretable and directly mapped to network indicators[1] (e.g., traffics and handovers).

To pursue this goal, in this paper, we explore for the first time the causal factor that directly impact the mobile voice quality but yet lacks attention—*the wireless network*. We observe that the major cause of the degradation of the mobile voice is nothing else but the signal propagation in wireless environment [18]. The network conditions such as network coverages, signal interferences, and mobility handovers together significantly affect the voice quality. Inspired by this, we design a wireless analytics algorithm, named CrowdMi, which models the mobile voice quality by mining various types of network indicators.

Manuscript received October 01, 2014; revised November 19, 2014; accepted December 12, 2014. Date of publication January 06, 2015; date of current version July 30, 2015.

Y. Ouyang is with Verizon Wireless, Basking Ridge, NJ 07920 USA, and also with the Institute of Data Science and Engineering, Columbia University, New York, NY 10027 USA (e-mail: yo2258@columbia.edu).

T. Yan is with NEC Laboratories America, Princeton, NJ 08540 USA (e-mail: yan@nec-labs.com).

G. Wang is with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: gwang@njit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JIOT.2014.2387771

[1] In this paper, we use the term network indicators to represent network performance and resource indicators.

CrowdMi respects POLQA as the standard for quantifying mobile voice quality. However, it does not directly measure the audio features. Instead, it builds a quantified causal relationship between the change of voice quality and the deviation of the network conditions, through large-scale data crowdsourcing from users in different network scenarios. With CrowdMi, in this paper we try to answer the following questions for mobile carriers: *How is the voice quality in your network? If not good, what causes that?*

In CrowdMi, to build the voice quality model, we make testing phones call each other and record voice audio clips and network indicators during the call. After data collection, we identify important RF features, and classify the data into different groups based on such features. Then, for each RF group, we design spatial silhouette distance (SSD) to select most relevant network indicators, based on which, we perform clustering to the data according to the selected network indicators that impose heavy impact on network performance. In each cluster, we then use the POLQA standard to analyze the audio features and compute the voice quality score. We design an adaptive LOESS algorithm to associate the network indicators to the computed scores by regressing the features that show highest correlation to voice quality. Finally, the quality model is built, where the selected network indicators are correlated and mapped to voice quality. The change of such features is the root cause of the deviation of the quality. After the model is built, CrowdMi no longer relies on POLQA. To assess the voice quality of a phone, it just collects the phone's network indicators, feeds into the model, and computes the estimated voice quality.

We follow client–server architecture to implement CrowdMi. The CrowdMi Client is implemented as an App in Android smartphones, which automatically collects user data in different locations and network scenarios through crowdsourcing, and sends back to the CrowdMi Server. The CrowdMi Server is deployed in our datacenter and runs our wireless analytics algorithm. It mines the collected data to build a model to model the mobile voice quality based on the collected network conditions. When the model is built and calibrated, it takes the realtime data collected from each of the CrowdMi Clients and calculates the mobile voice quality for the CrowdMi Client, which represent the current voice quality of the place where the CrowdMi Client locates. We deploy and conduct a pilot trial in the VoLTE network and crowdsource users in different geographic areas of the United States to study the network with different coverages. The trial shows that the CrowdMi does not require any additional hardware or human effort, and has very high model accuracy and strong diagnosability.

To summarize, the contribution of this paper is threefold.

1) We, for the first time, mine network indicators to achieve scalable and diagnosable mobile voice quality assessment.
2) We fully implement the CrowdMi, which runs the CrowdMi algorithm and collects data through user crowdsourcing.
3) We deploy our system and conduct a pilot trial in VoLTE networks, which shows the high usability of the system.

This paper is organized as follows. Section II introduces the existing related work. The main CrowdMi mining algorithm and its system implementation is described in Sections III and IV, respectively. Section V describes our pilot trial. Finally, we conclude this paper in Section VI.

II. Related Work and Background

In this section, we first introduce the state-of-the-art regarding to the voice quality assessment. After that, we describe the POLQA standard for voice quality assessment and discuss the motivation of our scheme.

A. Related Work

To the best of our knowledge, we are the first to address mobile voice quality assessment by mining wireless networks. There are not many comparable related works. In this section, we survey the closest works in voice quality assessment.

Research in assessing speech and voice quality mainly focuses on audio clips analysis [6]–[11], human voice modeling [12], [13], and language processing [14]–[16]. Berger *et al*. use short-term listening quality to evaluate the speech quality per call and calculate the mean opinion scores (MOS) to quantify the voice quality [6]. The edge-device is taken into consideration in [7] to measure the quality of voice-over-IP (VoIP) network. The intrusive speech quality model PESQ [19] standardized by ITU-T estimates the received quality of transmitted speech for the classical narrowband telephone bandwidth. It was extended in standardized ITU-T Rec. P862.2 [20] to model wideband transmissions. PESQ was recently replaced by POLQA [4] as the new standard for objective voice quality testing technology. Such objective quality testing was further studied in [9] and [11] to improve the flexibility of the evaluation, and other intrusive and nonintrusive speech quality assessment methods are described in [8] and [10]. Subjective voice quality assessment is conducted in [12] and [13]. A transmission planning tool, E-model, is designed in [21], and VQmon is provided in [22]. To further assess the quality of human voice, features in different languages and tones are extracted and modeled in [14]–[16]. However, while audio and language features may directly reflect voice quality, such evaluation is costly and cannot be directly applied in large-scale mobile network for network-wide assessment. Moreover, they do not consider the wireless network itself as a causal factor to the quality of the voice, and thus are unable to provide a guideline for network diagnose and optimization.

B. Preliminary of the POLQA Standard

POLQA is an ITU-T standard (ITU-T Rec. P.863) [4] that quantifies the quality of voice speech through audio signal analysis. It specially supports new types of speech codecs used in 3G and 4G LTE networks, and thus is widely adopted by mobile operators in estimating voice quality in 3G and VoLTE networks.

The key idea of POLQA is taking audio clips to be evaluated as the input and comparing it with the prerecorded reference audio signals to rate a degraded or processed speech signal in relation to the original signal. The difference between the two signals is counted as distortions. When the input clips end, the distorted speech files are scored from 1 to 5 based on MOS [23], and such score is the qualified assessment of the transmitted audio quality.

Practically, to evaluate the quality of mobile voice for cellphones, each test phone needs to connect to a POLQA box, which includes the POLQA assessment algorithm, microphone, audio recorder, playbacks, etc. Each POLQA box originates phone calls to other phones, plays the prerecorded the reference audio clips, and record the received audio signal (degraded). The recorded audio clips are then processed inside the box through the POLQA algorithm to calculate the quality score.

C. Why CrowdMi

As discussed in Section II-B, to analyze audio clips, the acoustic indicators are critical variables to quantify the quality of the audio samples. These acoustic indicators cannot be directly measured unless both of the following two conditions are satisfied. 1) A POLQA box including a stack of professional audio processing tools needs to be connected to each of the testing phones. 2) All the testing phones need to be paired in advance to launch/receive calls to/from each other. Such audio assessment involves much efforts from subject matter experts and huge hardware investments, and thus is only feasible for very small-scale testing in laboratory conditions. The overhead will go in exponential speed if we conduct large-scale evaluations, e.g., voice quality assessment for a carrier's network. Moreover, the audio feature analysis can only tell the quality of the voice, but is unable to identify its root cause, and thus is incapable of helping wireless operators, e.g., the mobile carrier, to diagnose and improve the network.

To overcome these issues, CrowdMi analyzes and mines the network indicators instead of acoustic signals to avoid huge evaluation overheads. It leverages the existing numerous mobile users for crowdsourcing realtime network data and survey the large-scale cellular network, without introducing extra hardware and human efforts. The crowdsourced network indicators are used to model the voice quality and learn the root cause.

III. CrowdMi Mining Algorithm

In this section, we first give an overview of our CrowdMi mining algorithm, and then describe the details of the CrowdMi algorithm.

A. CrowdMi Overview

CrowdMi consists of training phase and testing phase and can be installed as software in phones to perform voice quality assessment. In the training phase, to build the voice quality model, phones installed with CrowdMi make voice call to each other and collect the data, including audio clips and the network indicators during the call. After each call, the POLQA score of each voice clip record is computed using the method described in Section II-B, and the records are organized in a way such that at each time point, the audio quality score is associated with a set of network indicators and RF features.

With such data, based on domain knowledge and the recommendation by widely used standards [24], [25], we first classify the records into groups based on their RF quality. We identify two important RF indicators: 1) reference signal received power (RSRP) and 2) signal-to-interference-plus-noise ratio (SINR), which serve as features to classification. After that, each of the classified groups consists of records with a certain range of RF quality and the recorded network indicators. Then, inside each RF group, we perform clustering to cluster the records based on their network indicators. Before doing clustering, to reduce the overfitting, we want to only select important network indicators that are discriminative to separate records with good voice quality and ones with bad quality. To do so, we design SSD to measure the capability of each network indicator in differentiating voice qualities, and only select the features with large SSD value, e.g., SSD ≥ 0.7. The selected network indicators are treated as features, and we apply K-Medoids method to do the clustering. Such clustering selects the network indicators that impose heavy impact on network performance and group the data according to such features. Furthermore, in each cluster, we associate the selected network indicators to the computed POLQA scores by regressing the network indicators that have high correlation to voice quality. We propose adaptive local weight scatterplot smoothing (A-LOESS) regression, which improves original LOESS by adding adaptive window size to regress the features and compute the estimated voice quality score. Finally, the quality model is built, where in each cluster, selected network indicators are correlated and mapped to voice quality, and their change is the root cause of the voice quality deviation. After we build the voice quality model, we no longer rely on POLQA and have a full leverage of the model.

In the testing phase, to assess the voice quality for a phone, CrowdMi needs not to collect or analyze audio clips. Instead, it only collects network indicator and RF data of each phone. For each of the collected records, CrowdMi feeds it to the model to assign its RF quality group and network indicator cluster by measuring the similarity between the input records and the training model. After the assignment, it uses the trained A-LOESS model to compute the estimated voice quality for this record.

B. Classification on RF Quality

As mobile voice quality is mainly impacted by two facts, network coverage and interference, we follow the recommendation [26] by 3GPP TS 36.214 and 3GPP TS 36.133 to select two indicators: 1) RSRP and 2) SINR, to represent these two facts, and recommend their applicable ranges.

To be more specific, RSRP refers to the average power of resource elements that carry cell-specific reference signals over the entire bandwidth. RSRP is a direct cell signal strength indicator and thus is a representative indicator to denote the coverage strength. A strong coverage cannot ensure a good RF quality. A strong-covered area with high interference and noise may still has poor voice signal. Thus, SINR that reflects the interference and noise condition is used as a typical indicator to represent interference condition. Furthermore, domain experts [26] have proposed scales of LTE signal strength for

TABLE I
CLASSIFICATION ON RF QUALITY

Class no.	RSRP (dbm)	SINR (db)	Description
Class 1	≥−85	>15	Good Cov. and Low Intf.
Class 2	≥−85	≤15	Good Cov. and High Intf.
Class 3	(–105, –85)	>15	Median Cov. and Low Intf.
Class 4	(–105, –85)	≤15	Median Cov. and High Intf.
Class 5	≤−105	>15	Poor Cov. and Low Intf.
Class 6	≤−105	≤15	Poor Cov. and High Intf.

those signal indicators. Thus, there is no need to make a classification algorithm to train and derive each class, and we can simply use the scales to classify the RF quality. Table I is the proposed classification table based upon the scales of RSRP and SINR, respectively.

C. Feature Selection and Network Indicator Clustering

There are hundreds of kinds of network indicators in mobile networks, but only a few of them may impact the voice quality. In this section, we first design a feature selection method to select most relevant network indicators and then use them as features to perform clustering, which clusters the data to several groups based on the availability, sufficiency, and assignability of the network resources.

1) Selection of Network Indicators: We design SSD to select network indicators that are most discriminative to different voice qualities. In the training dataset, we first divide all the records into different quality groups according to their POLQA voice scores. Then, we calculate the SSD for each network indicator in each group, and use such value to determine the discrimination capability of the network indicator. More specifically, we follow ITU-T standard [27] to divide the mobile voice into four groups based on their POLQA score as follows: 1) C_1 : [0,2); 2) C_2 : [2,3); 3) C_3 : [3,4); and 4) C_4 : [4.0, 4.5].[2] Assume that each group C_k $(k = 1, 2, 3, 4)$ has n records and each record r_j^k has m network indicators. In each quality group C_k, for each network indicator point $R_{i,j}^k$ of each record r_j^k, we first compute the Euclidean distance (ED) to all the other points in the same group, and obtain the average intra-group ED $\mathrm{IntraED}_{i,j}^k$ for this feature point. Then, for this feature point, similarly, we compute its ED to feature points in all the other groups and calculate the average inter-group ED $\mathrm{InterED}_{i,j}^k$ for this network indicator point $R_{i,j}^k$. Following this way, we compute the average intra-group ED and average and inter-group ED for every feature point in the training records. After that, for each quality group, for each network indicator, we average out its intra-group ED over all the records inside the group to obtain its group-wise average intra-group ED $\mathrm{IntraED}_i^k$, and similarly, we obtain a group-wise average inter-group ED $\mathrm{InterED}_i^k$ for this indicator. Then, for each quality group C_i, the SSD for each indicator R_i^k is given by

$$\mathbb{S}_i^k = \frac{\mathrm{InterED}_i^k - \mathrm{IntraED}_i^k}{\max\{\mathrm{InterED}_i^k, \mathrm{IntraED}_i^k\}}. \tag{1}$$

For each network indicator R_i^k in a quality group C_k, we use tricube weight function to weight it as follows:

$$\mathbb{W}_i^k = \begin{cases} (1 - |\sum_{i,j=1}^{4}(R_i^k - R_j^k)|^3)^3, & \text{if } |\sum_{i,j=1}^{4}(R_i^k - R_j^k)| < 1 \\ 0, \quad \text{if } |\sum_{i,j=1}^{4}(R_i^k - R_j^k)| \geq 1. \end{cases} \tag{2}$$

Finally, we obtain the SSD for each network indicator R_i over all the RF groups

$$\mathbb{S}_i = \mathbb{W}_i^k \times \mathbb{S}_i^k. \tag{3}$$

After we obtain the SSD for all the network indicators, we select indicators with $\mathbb{S}_i \geq 0.7$ as the features to perform clustering, considering they are the discriminative features and are highly correlated to the voice quality.

2) Clustering: After selecting the discriminative features, we perform clustering to all the training records using the selected features. We use K-Medoids clustering algorithm by imposing a new converging rule to identify the best k. The reason to adopt K-Medoids rather than other distance-based clustering algorithm is that network indicator consumption show high deviation due to peak time and rush hours. Hence the network indicators, containing many spikes and outliers, may be diluted in K-Means and the other similar algorithms. To choose the optimal number of clusters k, we define an upper bound of cluster number u based on domain experience. We iterate k from 2 to u and perform K-Medoids clustering at each iteration. The optimal k is selected, such that the intra-cluster error is minimized and the inter-cluster distance is maximized as follows:

$$\begin{cases} 0.7 \leq \dfrac{\mathrm{IntraSumOfError}_{k+1}}{\mathrm{IntraSumOfError}_k} \leq 1 \\ 0.7 \leq \dfrac{\mathrm{IntraSumOfError}_{k+2}}{\mathrm{IntraSumOfError}_{k+1}} \leq 1 \\ 0.7 \leq \dfrac{\mathrm{IntraSumOfError}_{k+3}}{\mathrm{IntraSumOfError}_{k+2}} \leq 1. \end{cases} \tag{4}$$

D. A-LOESS Feature Regression Based on POLQA Scores

After clustering all the records into different clusters, we conduct a regression to regress POLQA for each cluster. We propose A-LOESS to regress POLQA scores based on the network indicators selected in Section III-C1. A-LOESS improves LOESS [28] by adaptively computing a proper window size during the regression, instead of the fixed window size in the original LOESS. More specifically, we pack the POLQA scores into different bins, and dynamically adjust window size for each local set by the distribution density of each bin. Based on domain experience in voice assessment in POLQA, we set nine bins according to the POLQA scale: $\mathrm{bin}_0 = [0, 0.5]$, $\mathrm{bin}_1 = (0.5, 1)$, $\mathrm{bin}_2 = [1, 1.5], \ldots, \mathrm{bin}_8 = (4.5, 5]$. We set an initial window width to $1/100$ of range of sample points, and plot the scatterplot of all measured POLQA scores in an ascending order. Let $f(x)$ denote the scatterplot function, where x

[2] For SWB mode, C_4 needs to be [4.0, 4.75], as the maximum POLQA score in such mode is 4.75.

is from 1 to the number of POLQA sample points. First, for each bin bin_a, we compute its distribution density by integrating the value of the scatterplot function in its range as follows:

$$y_a = \int_{f^{-1}(0.5a)}^{f^{-1}(0.5a+0.5)} f(x)\,dx, \quad i = 0, \ldots, 8. \tag{5}$$

After that, we sort y_a in ascending order. Let $S(y_a)_{\min}$ represent the bin with minimum y_a, $S(y_a)_{\text{med}}$ represent the bin that has the median value of y_a, and $S(y_a)_{\max}$ represent the bin with maximum y_a. We dynamically calculate the window size by the sorting results, as follows:

$$\text{win\_size} = \begin{cases} \dfrac{0.5 + 0.125 \cdot S}{100} \cdot N, & \text{if } S = 0, \ldots, 4 \\ \dfrac{1 + 0.25 \cdot (S - 4)}{100} \cdot N, & \text{if } S = 5, \ldots, 8. \end{cases} \tag{6}$$

Finally, we use the adaptive window size calculated by (6) to perform LOESS regression to the POLQA score based on the selected features.[3]

IV. CrowdMi System

CrowdMi consists of two major components, CrowdMi Client and CrowdMi Server. It operates in two phases: 1) training phase that collects data and build voice quality model and 2) test phase that crowdsources user data to assess voice quality, as illustrated in Fig. 1. The CrowdMi Client is implemented as an App in Android smartphones, and its main functionality is to collect user data in different locations and network scenarios through crowdsourcing, and send back to the CrowdMi Server. The CrowdMi Server runs our wireless analytics algorithm. In the training phase, it mines the collected data to build a model to model the mobile voice quality based on the collected network conditions. When the model is built and calibrated, in the testing phase, it takes the realtime data collected from the each of the CrowdMi Clients and calculates the mobile voice quality for the CrowdMi Client, which is the current voice quality of the place where the CrowdMi Client locates.

A. *CrowdMi Client*

The CrowdMi Client in smartphones automatically monitors the network conditions of the phones and collects the data. In the training phase, each of such phones is operated by test engineers and connected with a POLQA box, the standardized voice quality measuring system. The POLQA box includes several prerecorded audio clips of reference speech, and a standardized objective voice quality measurement system that takes input voice clips, compares such speech with the reference speech, and calculates the quality of the voice. When training phase starts, phones with the CrowdMi Client call each other, play audio clips generated by the POLQA box, record the audio clips received from the other phone, and at the same time record the network conditions of the phone during the call. After each call ends, each POLQA box calculates the quality score of the recorded audio clips, and the client uploads the score and network indicators to the CrowdMi Server. The CrowdMi Server uses such data to build a model for voice quality assessment. In the testing phase, the CrowdMi Client leverages the existing numerous mobile users and is installed in their phones. The phone needs not to connect to POLQA box and only runs the CrowdMi Client. The client does not make phone calls, and runs in the background just to collect network indicators of the phone. It sends data back to the CrowdMi Server periodically, reporting the network conditions of the phone in different locations.

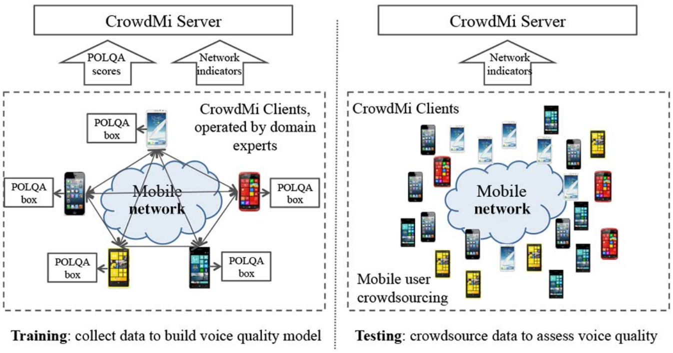

CrowdMi Server
POLQA scores
Network indicators
CrowdMi Clients, operated by domain experts
POLQA box
POLQA box
POLQA box
POLQA box
POLQA box
Mobile network
Training: collect data to build voice quality model
CrowdMi Server
Network indicators
CrowdMi Clients
Mobile network
Mobile user crowdsourcing
Testing: crowdsource data to assess voice quality

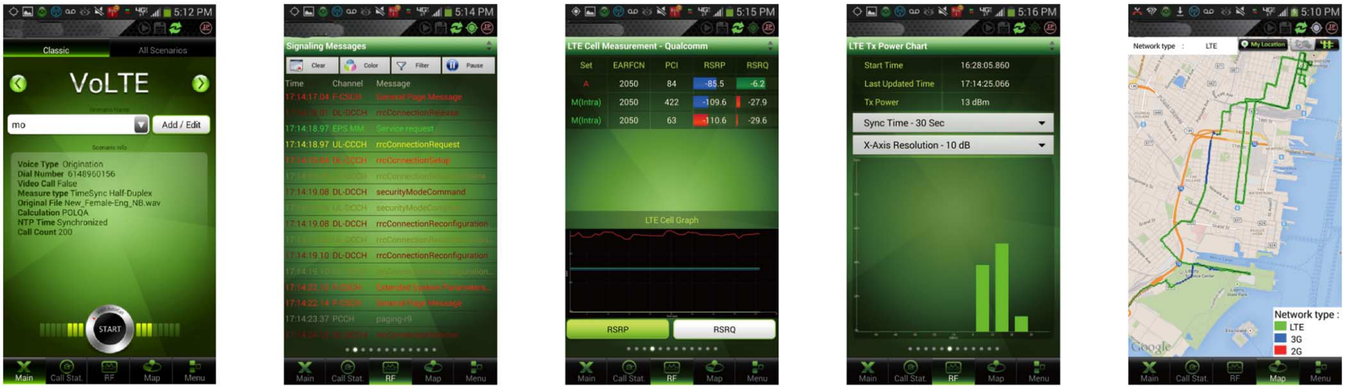
Fig. 1. Architecture of the CrowdMi system.

The screenshots of the CrowdMi App running in training phase in VoLTE scenario is shown in Fig. 2. To help domain engineers diagnose the network issues, the App also include rich log information and the visualization of assessment results sent back from the CrowdMi Server, such as KPIs, quality assessment scores, and location traces. Such information can be displayed in realtime and is shown in various types, which drastically facilitates the voice assessment of the network. Fig. 3 shows an example of a CrowdMi Client in a testing vehicle during training phase. It is worth to note that the POLQA box needs not to be connected, and the visualization features can be turned OFF when doing the large-scale crowdsourcing in the testing phase. The App runs silently in the background and does not disrupt any other cellphone usages.

B. *CrowdMi Server*

The CrowdMi Server builds a voice quality model and assesses the voice quality of the cellular networks in different locations and coverage conditions. In the training phase, the server collects data from clients, and runs our CrowdMi mining algorithm to model the mobile voice quality using the received voice quality scores and network indicators. After the model is built, it is stored in the server. In the testing phase, for each client, the server periodically estimates the voice quality using the computed model, and such estimation is the voice quality assessment of the network in the client's location.

[3] We omit the description of the well-known LOESS algorithm. Please refer to [28] for details.



(a)

VoLTE

Scenario Name: mo

Scenario Info

Voice Type Origination
Dial Number 6148960156
Video Call False
Measure type TimeSync Half-Duplex
Original File New_Female-Eng_NB.wav
Calculation POLQA
NTP Time Synchronized
Call Count 200

(b)

Signaling Messages

Time	Channel	Message
17:14:17.04	F-CSCH	General Page Message
17:14:18.01	DL-DCCH	rrcConnectionRelease
17:14:18.97	EPS MM	Service request
17:14:18.97	UL-CCCH	rrcConnectionRequest
17:14:18.94	DL-CCCH	rrcConnectionSetup
[illegible]	UL-DCCH	rrcConnectionSetupComplete
17:14:19.08	DL-DCCH	securityModeCommand
[illegible]	UL-DCCH	securityModeComplete
17:14:19.08	DL-DCCH	rrcConnectionReconfiguration
[illegible]	UL-DCCH	rrcConnectionReconfiguration...
17:14:19.10	DL-DCCH	rrcConnectionReconfiguration
17:14:19.10	[illegible]	...onfiguration...
17:14:22.12	F-CSCH	Extended System Parameters...
17:14:22.14	F-CSCH	General Page Message
17:14:23.37	PCCH	paging-r9
17:14:24.12	DL-DCCH	rrcConnectionRelease

(c)

LTE Cell Measurement - Qualcomm

Set	EARFCN	PCI	RSRP	RSRQ
A	2050	84	-85.5	-6.2
M(Intra)	2050	422	-109.6	-27.9
M(Intra)	2050	63	-110.6	-29.6

LTE Cell Graph

(d)

LTE Tx Power Chart

Start Time	16:28:05.860
Last Updated Time	17:14:25.066
Tx Power	13 dBm

Sync Time - 30 Sec

X-Axis Resolution - 10 dB

(e)

Network type : LTE

Fig. 2. Screenshots of the CrowdMi App running in LTE network scenario. (a) Front page of the app. (b) Log information during data collection. (c) Visualization of network conditions. (d) Visualization of network conditions. (e) Visualization of moving traces during data collection.

Fig. 3. Example scenario of a CrowdMi Client smartphone in training phase.

V. Pilot Trial

A. Setting

To verify our CrowdMi system, we conduct a pilot trial in a VoLTE network of different geographic areas with various network qualities in a major network carrier. Our objective in conducting this evaluation study is to test accuracy of the voice quality model and evaluate the diagnosability of the system for finding relevant network indicators to the voice quality. The trial lasts for 9 months, from December 2013 to August 2014. During the 9 months, we install CrowdMi Clients to 50 smartphones with Android 4.3 System that supports VoLTE functionality. The clients measure all the needed network/RF/device performance indicators, and collect and upload test logs on a rotation basis.

To collect voice data, we select the prerecorded List-11 Harvard Sentences of Female American English voices [29], each with 10 s length, as the audio input for the POLQA box. All the testing phones are in time synchronization half-duplex mode. When a phone calls another one, and plays the audio clip, the receiver starts to compute POLQA score by comparing the received audio signal against the reference audio signal, and at the same time, it starts to play the same audio clip back to the caller.

TABLE II
Selected Features

Features	SSD
MAC.DL.Throughput	0.8430
PDSCH.Throughput	0.8214
RLC.DL.Throughput	0.8186
RTP.Rx.Throughput	0.8057
PDCP.DL.Throughput	0.7934
RTP.Tx.Throughput	0.7928
RTP.Rx.Delay	0.7412
RTP.Rx.Jitter	0.7103
Handover.Happening	0.6974

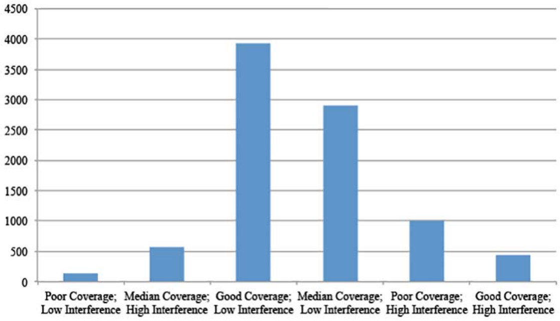


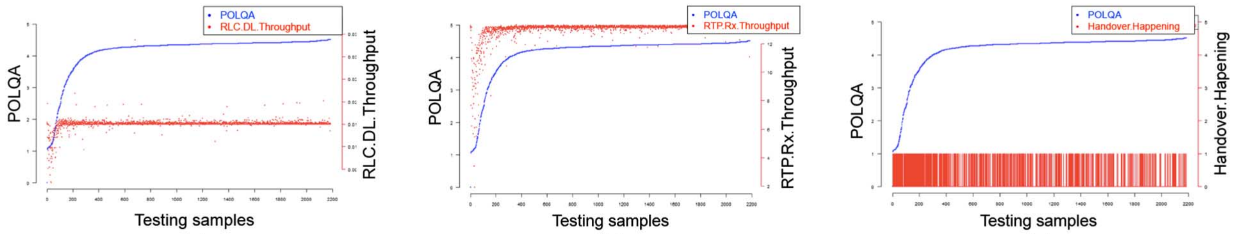
Fig. 4. Record distribution in RF groups.

B. Data

Most of the data collections are performed by drive tests considering that mobility is an important factor to voice quality. Among all the data logs, 77% are drive tests and 23% are stationary tests. We randomly select wireless environment for each test case. In this way, we generate POLQA records in diverse wireless environments with different qualities of coverage and interference. In total, we collected 317 logs of POLQA test cases, where 299 are valid and 18 are error logs and thus discarded. The valid logs consist of 8987 POLQA voice records. According to Table I, all the records are classified into six groups by the measured RSRP and SINR values. Fig. 4 shows the distribution.

C. Feature Selection

We apply SSD to select discriminative network indicators that have high impact on the voice quality. Table II shows the top nine selected ones. From this table, we can see that majority



(a) (b) (c)

Fig. 5. Selected features (blue) plotted together with the computed POLQA scores (blue). (a) RLC.DL.Throughput. (b) RTP.Rx.Throughput. (c) Handover.Happening.

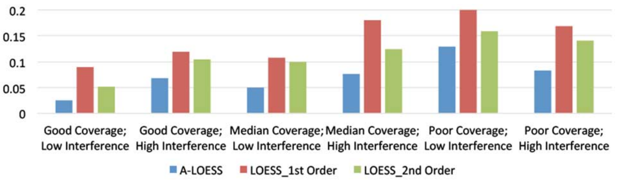


Fig. 6. MAPE in training dataset.

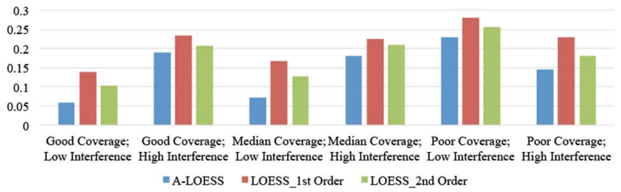


Fig. 7. MAPE in test dataset.

of the selected indicators are related to throughput and audio transmission, which is under our expectation.

To show the high correlation of the selected features to the voice quality, we select RLC.DL.Throughput (the throughput of the RLC layer), RTP.Rx.Throughput (the realtime audio transmission throughput), and Handover.Happening (whether a handover happens or not), from Table II, and plot them with the computed POLQA score in Fig. 5. As expected, the POLQA score is high when the throughput indicators are high as shown in Fig. 5(a) and (b), and the score is low when handover happens frequently as shown in Fig. 5(c). The indicators are strongly correlated to the POLQA scores.

D. Accuracy of the CrowdMi Model

In CrowdMi, after feature selection, we use the selected features to perform K-Medoids clustering. In each cluster, we design A-LOESS algorithm to regress the features, and estimate the POLQA score based on the network indicators.

To evaluate the accuracy of our A-LOESS algorithm, we use 75% of the data as training dataset and the rest 25% as testing dataset. We compare our designed A-LOESS model with two other models, original LOESS with first-order smoothing and original LOESS with second-order smoothing. We use mean absolute percentage error (MAPE) to measure the error of the models as follows:

$$e = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{S_i^{\text{POLQA}} - S_i^{\text{Est}}}{S_i^{\text{POLQA}}}\right| \quad (7)$$

where S_i^{POLQA} is the true POLQA score and S_i^{Est} is the estimated POLQA score by models for the ith record, respectively.

The training and test MAPE of all the compared models in each RF group is shown in Figs. 6 and 7, respectively. From Fig. 6, we can see that, all the training MAPEs of A-LOESS are lower than 10% except for group "poor coverage and low interference." Actually, the high MAPE in this group is not caused by our model, but insufficient records collected from the trial as shown in Fig. 4. We can overcome this issue by conducting a few additional tests in environments of such RF group. Comparing with other two methods, the error of A-LOESS in training set is the lowest for all the six groups. Similarly, from Fig. 7, we can see that A-LOESS performs better than the other two methods in test datasets. Overall, the MAPE of our scheme is maintained at a very low level and is lower than the compared schemes, which indicates great model accuracy of our CrowdMi system. Moreover, the reliability between training and test set is robust, as the difference of MAPE between training and test is small, which is no larger than 12.58%, also from group "poor coverage low interference." This shows CrowdMi is a valid approach that can be applied for POLQA assessment in LTE networks.

VI. Conclusion

In this paper, we design CrowdMi, a wireless analytics system, to assess mobile voice quality by crowdsourcing and analyzing the network conditions of cellphones. CrowdMi mines hundreds of network and RF indicators to build a causal relationship between voice quality and network conditions, and carefully calibrates the model according to the widely accepted POLQA voice assessment standard. It avoids the costly analysis of audio clips and achieves high scalability and diagnosability. We fully implement the CrowdMi, including a server running the CrowdMi mining algorithm, and clients installed in Android smartphones that collect data through user crowdsourcing. We deploy our system and conduct a pilot trial in VoLTE networks in the United States, which shows the high usability of the system.

Acknowledgment

The authors would like to thank Prof. M. Hosein Fallah, Stevens Institute of Technology, for the stimulating discussion.

References

[1] Business Insider. (2013, Jun.). *How Much Time Do We Really Spend on Our Smartphones Every Day?* [Online]. Available: http://www.businessinsider.com.au/how-much-time-do-we-spend-on-smartphones-2013-6

[2] Evaluation Engineering. *Factors to Consumer Satisfaction* [Online]. Available: http://www.evaluationengineering.com/articles/200806/mobility-performance-testing.php

[3] *QoS in Cellular Networks*, Washington Univ., St. Louis, MO, USA [Online]. Available: http://www.cs.wustl.edu/jain/cse574-06/ftp/cellular_qos/index.html

[4] ITU. *Perceptual Objective Listening Quality Assessment* [Online]. Available: http://www.itu.int/rec/T-REC-P.863

[5] T. Falk and W. Chan, "Hybrid signal-and-link-parametric speech quality measurement for VoIP communications," *IEEE Trans. Audio Speech Language Process.*, vol. 16, no. 8, pp. 1579–1589, Nov. 2008.

[6] J. Berger *et al.*, "Estimation of quality per call in modelled telephone conversations," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 4809–4812.

[7] S. Broom, "VoIP quality assessment: Taking account of the edge-device," *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 6, pp. 1977–1983, Nov. 2006.

[8] N. Ct, V. Koehl, V. Gautier-Turbin, A. Raake, and S. Mller, "An intrusive super-wideband speech quality model: Dial," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech'10)*, Makuhari, Japan, 2010, pp. 1317–1320.

[9] W. Zhang, Y. Chang, Y. Liu, and L. Xiao, "A new method of objective speech quality assessment in communication system," *J. Multimedia*, vol. 8, no. 3, pp. 291–298, Jun. 2013.

[10] Q. Li, Y. Fang, W. Lin, and D. Thalmann, "Non-intrusive quality assessment for enhanced speech signals based on spectro-temporal features," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 2014, pp. 1–6.

[11] P. Bauer, C. Guillaumea, W. Tirry, and T. Fingsheidt, "On speech quality assessment of artificial bandwidth extension," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2014, pp. 6082–6086.

[12] P. Reichl, S. Egger, R. Schatz, and A. DAlconzo, "The logarithmic nature of QoE and the role of the Weber-Fechner law in QoE assessment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2010, pp. 1–5.

[13] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.

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

[15] M. Gueguin, R. LeBouquin-Jeannes, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," *EURASIP J. Adv. Signal Process.*, vol. 2008, pp. 1–15, 2008, Article ID 185248.

[16] B. Weiss, S. Moller, A. Raake, J. Berger, and R. Ullmann, "Modeling conversational quality for time-varying transmission characteristics," *Acta Acoust. United Acoust.*, vol. 95, no. 6, pp. 1140–1151, 2008.

[17] E. Uzoamaka, "Validating perceptual objective listening quality assessment methods on the tonal language Igbo validating perceptual objective listening," M.S. thesis, Dept. Comput. Sci., Delft Univ. Technol., Delft, The Netherlands, 2009.

[18] G. L. Stuber, *Principles of Mobile Communication*, vol. XXIII, 3rd ed. New York, NY, USA: Springer, 2012, 830pp.

[19] ITU. *Perceptual Evaluation of Speech Quality (PESQ)* [Online]. Available: http://www.itu.int/rec/T-REC-P.862/en

[20] ITU. *wB-PESQ* [Online]. Available: http://www.itu.int/rec/T-REC-P.862.2

[21] ITU. *e-Model* [Online]. Available: http://www.itu.int/rec/T-REC-G.107

[22] TelChemy. *vQmon* [Online]. Available: https://www.telchemy.com/vqmon.php

[23] ITU. *Mean Opinion Score* [Online]. Available: http://www.itu.int/rec/T-REC-P.800-199608-I/en

[24] ITU. *3GPP TS 36.214: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements* [Online]. Available: http://www.itu.int/rec/T-REC-P.862.2

[25] *3GPP TS 36.133: Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for Support of Radio Resource Management* [Online]. Available: http://www.itu.int/rec/T-REC-P.862.2

[26] M. Sauter, *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*. Hoboken, NJ, USA: Wiley, 2010.

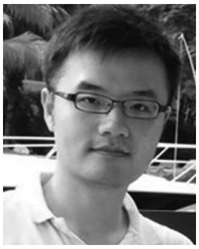
[27] ITU. *ITU-T Recommendation P.800 Methods for Subjective Determination of Transmission Quality* [Online]. Available: http://www.itu.int/rec/T-REC-P.800-199608-I/en

[28] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *J. Amer. Stat. Assoc.*, vol. 74, no. 368, pp. 829–836, 1979.

[29] *Harvard Sentences*, Columbia Univ. [Online]. Avilable: http://www.cs.columbia.edu/hgs/audio/harvard.html

Ye Ouyang received the B.E. degree from Southeast University, Nanjing, China, the M.S. degree from Tufts University, Medford, MA, USA, and the Ph.D. degree from the Stevens Institute of Technology, Hoboken, NJ, USA.

He is a Distinguished Staff Scientist with Verizon Wireless, Basking Ridge, NJ, USA. In 2012, he was, as Co-Principal, awarded telecom research funding by the White House, Office of Science and Technology, Washington, DC, USA. His research lies in big data analytics for wireless networks, with a focus on 2G/3G/4G LTE network performance, network capacity, traffic patterns, user behaviors, and network and device service quality.



Tan Yan received the B.E. degree in information science and technology from Southeast University, Nanjing, China, in 2007, the M.S. degree in electrical and computer engineering and Ph.D. degree in computer science from the New Jersey Institute of Technology (NJIT), Newark, NJ, USA, in 2009 and 2014, and respectively.

He joined NEC Laboratories America, Princeton, NJ, USA, in 2014. His research interests include network analytics, time-series mining, mobile *ad hoc* data management and dissemination, and graph theory.

Guiling Wang received the B.S. degree in software from Nankai University, Tianjin, China, and the Ph.D. degree in computer science and engineering from Pennsylvania State University, State College, PA, USA, in 2006.

She joined the New Jersey Institute of Technology (NJIT), Newark, NJ, USA, in 2006, and became an Associate Professor with tenure in 2011.