An Affect-based Approach for QoE evaluation in VoIP Systems

Abhishek Bhattacharya, Zhenyu Yang and Deng Pan
School of Computing and Information Sciences, Florida International University
Email: {abhat002, yangz, pand}@cis.fiu.edu

Abstract—The success of VoIP systems such as Skype, Google Talk, MSN Messenger, etc. have inspired the migration of voice communication from the Public Switched Telephone Network (PSTN) to the Internet. The inherent challenges for designing VoIP systems are due to the particular characteristics of voice communication such as low volume, burstiness and stringent delay/loss requirements which is different from the best-effort data delivery model of Internet. Effective evaluation frameworks are important for further refinement of these systems by adopting user feedback based requirements. Most of the VoIP evaluation models are system-centric (Quality of Service or QoS-based), which questioned us to explore for an user-centric (Quality of Experience or QoE-based) approach as a step towards the human-centric paradigm. We research an affect-based QoE evaluation framework which tries to capture user's perception while they are using VoIP applications along multiple dimensions. Our experimental methodology is illustrated in depth and we provide detailed results.

I. INTRODUCTION

The tremendous popularity of VoIP systems in the last few years providing real-time voice communication over any IP network (public or private) have made a significant impact on the multi-billion dollar telecommunication industry. As more and more service providers are developing VoIP products over increasingly heterogenous networks (such as ip, wireless, etc.), the evaluation models for effectively determining system performance are becoming critically important. Though one of the main characteristics of VoIP system is user-interactivity which is demonstrated by the multi-party collaboration through the audio channel, but still the evaluation frameworks remain very much system/network-centric. Human-centric computing (HCC) approaches are recently gaining importance in the multimedia research community with a paradigm shift from system/network-centred to user-centred design and evaluation methods. The gap between system/network and human-centric evaluation is responsible for the peripheral drawbacks of QoS-tuned systems which fail to capture the user’s perspective. Quality of Experience or QoE concept is essentially developed to address this issue but the present knowledge of QoE is still very blurred with varied interpretations encompassing the domains of psychology, cognitive science, sociology, and information technology. An influential work on QoE modeling approach was proposed by [15], where QoE is constructed as a causal chain of environmental influences→cognitive perceptions→behavioral consequences with QoS and QoE being viewed as distinct components in contrast to the traditional belief of thinking QoE as an extension or a subset of QoS.

Methods for measuring the quality of audio/speech applications are broadly classified as subjective or objective. Subjective methods are also intrusive since it requires user's input in the evaluation process. The most common way of measuring subjective quality involves in the utilization of the user’s cognitive resource by rating the test sample in a scale from 1 to 5 generally known as MOS (Mean Opinion Score). The major drawbacks of subjective quality evaluations are: (1) expensive and slow process involving direct involvement of user, (2) user’s feeling are subject to various effects such as personal bias and memory effect. OneClick [8] proposed a very interesting idea for subjective evaluation where the user needs to click whenever he/she feels dissatisfied with the quality. OneClick solved two significant problems prevalent in subjective evaluations: (1) a simple dichotomous decision rather than a multiple-choice decision, (2) time-varying quality of network applications where users are limted by their finite memory and subjected to recency effect. But, OneClick is still intrusive where the users need to be always alert for clicking whenever he/she is dissatisfied which involves cognitive overhead.

On the other hand, objective methods are nonintrusive and does not require user’s involvement, but they are based on well-validated data derived from subjective evaluations. Objective evaluations can be signal-based models or parametric models (considers network/application QoS factors). The current standard ITU-T P.862 also known as Perceptual Evaluation of Speech Quality or PESQ represents the state-of-art double-ended algorithm. Some variants of signal-based models are also no-reference or single-ended and they can estimate quality by processing only the degraded received speech signal represented by the current state-of-art ITU-T standard P.563 for single-ended quality measurement. The weakness of objective measures are: (1) cannot assess all the QoE dimensions since no explicit user involvement, (2) factors such as burstiness, path delay variability are not considered, (3) does not model the conversational dynamics i.e. different switching speed with various talk/silent spurt durations, and thus different sensitivities to delay. E-model is the ITU-T standard for quality evaluation of network applications that uses transmission parameters to predict the subjective quality of packetized voice. Thus, we observe that both the subjective and objective evaluation models possess certain inherent
drawbacks, which are not perfect solutions covering all the quality artifacts, and so a tradeoff relation exist between them in the real scenario.

Affective Computing deals with automated analysis of human affective behavior which has attracted increasing attention from researchers due to its multidisciplinary nature spanning psychology, computer science, linguistics, neuroscience, etc. Affect-aware applications are gaining acceptance due to its human-centric attribute which can adapt to the changing emotion of the user and effective in various systems such as user interface, gaming, mental health, learning technologies, customer services, intelligent automobile, entertainment industry, etc. Emotion is closely related to decision-making and thus plays a significant role in the action of human beings as shown in research by psychologists and neuroscientists [11]. An automatic dialog system is reported in [11] which has the ability of responding to callers according to the detected emotional state or they can pass control over to human operators. Extensive studies have been performed for automated emotion recognition in the context of human-computer interaction by exploring multiple input modalities such as facial expression, speech, body gestures, physiological signals, posture, neuroimaging, etc. for extracting implicit feedback using signal processing, linguistic analysis, text processing and various other techniques [16]. Users interact among each other with intentions, motivations and feelings besides real-life problems and information objects, which are all critical aspects of cognition and decision-making [5]. The influence of affect on user motivation, search strategies, performance, satisfaction are strong implicit indicators which provides feedback cues for evaluating web/multimedia search relevance [5]. Based on the above evidences, we hypothesize QoE or perception of quality by user as a implicit decision-making phenomenon related to human cognitive state that derives from the sensory channels and is highly correlated to user affective behavior. According to the best of our knowledge, affective behavior study in the context of remote human-human interaction integrated through a network channel and its dynamics with delay, loss-rate, bandwidth fluctuations is a relatively new and unexplored area.

We investigate various affective cues from users interacting in a VoIP system which are extracted in an implicit manner and its association with evaluation of quality of experience.

Our presented work in this paper delves into the role of affect in the VoIP quality experience evaluation and the potential impact of system/network/cognitive/behavioral fluctuation on user’s emotional behavior. However, we do not assume anything about the details of the relationship between user’s affective responses and QoE evaluation; we systematically build our ground truth from the data using training, classification and pattern recognition methods. We present detailed experimental methodology with real user studies and discuss outcome of the results. Section II discusses the affective models and the classification techniques that were employed. Section III provides the experimental design of our study and its subcomponents with implications of the result on the prediction of quality of experience with the use of affective feedback.

Finally, Section IV concludes the paper and provide future directions.

II. AFFECT BASED MODELING

Our choice of affect-based approach for QoE evaluation is linked to the fact that human emotions have a high degree of semblance to subjective user perception and also plays a decisive role in the cognitive decision making process of quality evaluation. We examine the following research hypothesis:

- **H**: Users’ affective responses as determined from vocal expression processing, will vary across the subjective perception of the user which influences the cognitive decision making process of quality evaluation over dynamic network scenarios.

In Section II-A we present the overview of our affective framework for quality evaluation. Section II-A1 describes the acoustic module of the framework succeeded by the linguistic component in Section II-A2. Next, we examine the classifier unit of the framework in Section II-B with various schemes for combining classifiers and finally some details about the output labels in Section II-C.

A. Affective Framework

The most basic affective cues that are extracted from voice signals for the purpose of emotion detection are known as acoustic features or Low-Level Descriptors (LLD). A comprehensive survey of various acoustic components for automatic emotion detection is studied in [16]. Most of the automatic emotion detection systems attempt to classify human emotion according to the six basic expressions: happiness, sadness, anger, fear, surprise, and disgust. Automatic detection of the six basic emotions in posed, controlled audio can be performed with reasonably high accuracy but detecting these expressions or any expression of affective behavior in less constrained or spontaneous settings is a very challenging problem. It is found that the closer we get to a realistic scenario with spontaneous human behavior, the less reliable is prosody as an indicator of the speaker’s emotional state. So, it is derived that prosodic features are required to be combined with other knowledge sources such as linguistic information to improve recognition accuracy [16]. Based on the above fact, we conceptualize our affective framework for QoE evaluation with the following modules: acoustic and linguistic. We introduce each of them in the following sections.

1) **Acoustic Features**: In this module we consider 22 different acoustic attributes related to segmental and suprasegmental information speech signals. These features were extracted by turn-level statistics and certain statistical functionals and transformations are applied corresponding to fundamental frequency (F0), energy, duration, and the first and second formant frequencies.

*Fundamental Frequency (F0)*: mean, median, standard deviation, maximum, minimum, range (max-min) and linear regression coefficient.

*Energy*: mean, median, standard deviation, maximum, minimum, range, and linear regression coefficient.
• Duration: speech-rate, duration of the longest voiced speech and ratio of voiced and unvoiced region.
• Formants: first and second formant frequencies (F1, F2), their bandwidth (BW1, BW2), and their mean.

The average length of voiced portion of speech was utilized for the speech-rate calculation. The voiced portion of speech was also used for calculating the linear regression coefficients for F0. Thus, the base feature set is composed of 22 (Base) different acoustic attributes extracted from the speaker’s voice signal. We employ the leave-one-out method for feature selection and using nearest neighborhood rule for estimating accuracy as described in [11]. We generated two variants of feature set, one had 10 (f10) best features and the other had 15 (f15) best features. We also obtain another version of feature set by operating Principal Component Analysis (PCA) on the base set.

2) Linguistic Features: Linguistic Features are language related information which exploits the fact that people tend to use specific word choices for expressing their emotions under different situations. For example, certain word utterances such as “no” or swear words have a high probability to be associated with negative subjective perception. We adopt the emotional salience concept described in [11] for modeling the linguistic feature module. An emotional salient word with respect to a quality category is one which appears more often in one category than in other. The feature dimensions chosen for this set are the difference of activations [11] i.e., \(a_0 - a_1\). We slightly modify this value to suit our framework where we have three output labels (will be discussed in detail later) i.e., \(a_0 - a_1, a_1 - a_2, \) and \(a_2 - a_0\). The word utterances were generated from speech signals by applying current state-of-art Automatic Speech Recognizer or ASR. We also take repetition metric for modeling linguistic features since it was found to be the most important indicator of trouble in communication [6].

We derive an empirical approach by considering 1 to 5-word repetitions only for keeping the computational overhead and linguistic information within satisfiable limits. We formulate the dimensions of this feature set as: number of 1-word repetition, number of 2-word repetitions, and likewise.

B. Classifier & Combinations

We extract a set of diverse features from multiple information sources and then perform discriminant analysis, using a classification method of Support Vector Machines (SVM). For SVM classifier, we trained our models using radial basis function (RBF) kernel, which among the four basic SVM kernels (linear, polynomial, radial basis function and sigmoid) are preferable due to its lesser numerical difficulties and a limited number of parameters. We trained three different derivatives of SVM models: (1) SVM–Plain: plain models using separate training and testing sets, (2) SVM–5CV: 5-fold cross-validation, and (3) SVM–5WC; SVM–10WC: 2-layer hierarchical SVM models with 5 and 10 weak classifiers i.e., training 5/10 classifiers on a different subset of the training set and the output of each classifier is used to train a meta-classifier.

Since we have incorporated multiple information sources (i.e. acoustic, linguistic) in our framework, so we need to consider the problem of combining these sources or classifiers (one classifier for each information source) in an effective fashion to generate a single predicted output label. We administer a simple average of the outputs from each of the classifiers since it achieves good performance inspite of its simplicity. Let \(x_n\) and \(y_n\) denote input and output for a classifier that processes different source of information where the total number of classifiers are \(n = 1, ...N\), while \(y\) be the final output and \(x\) represents a feature set with the whole information, then the final output of the overall classifiers is given as follows:

\[
y(x) = \frac{1}{N} \sum_{i=1}^{N} y_n(x_n)
\]

C. Output Labels

The output labels convey the subjective perception of the VoIP users captured through a rating scale to describe the user experience or QoE. However, we refrain from using the 5-point scale due to the obvious drawbacks as discussed in Section I. We employ a trichotomous or 3-point scale by augmenting with an average label which basically suffices for the situation when an user is unable to make a decision between ‘good’ and ‘bad’. During the experimental procedure, we asked each user about their feeling between a 2-point and a 3-point scale, and 80% (12 out of 15 participants) felt the need for an ‘average’ label when they are faced with a cognitive dilemma between ‘good’ and ‘bad’. The complete schematic diagram of our affective framework is shown in Figure 1.

III. EXPERIMENTAL METHODOLOGY AND RESULTS

We perform experimental evaluation for validating our hypothesis of QoE evaluation framework based on affective feedback of the speaker.

A. Experiment Setup

We installed two desktop computers with Intel i686 Core 2 Quad CPU 32-bit (2.39GHz) and 2GB RAM running Linux.
kernel ver 2.6.31.5 at each end of the VoIP setup. We also configured another similar computer, as a layer-2 bridge between the two ends with two network interface cards and applied the Linux bretl utility for a logical connection between both ends. For simulating long-distance network connection, we instrument the traffic flowing between the two ends by applying dummynet [12]. We used the PortAudio [4] library for the purpose of audio recording from the microphone into wav files. For the purpose of VoIP software, we employed PJ-SIP [3] which is an open-source, comprehensive, high performance, extremely portable system with a small memory footprint.

B. Network Environment Modeling

In order to validate our hypothesis, we need to instrument the network QoS parameters i.e., delay, bandwidth, loss-rate and to study its reaction on affective feedback of the user which will essentially help to predict the QoE. Each network setting is denoted as a QoS tuple of (delay, bandwidth, loss-rate). The next step involves conducting multiple test sessions with different QoS tuples, but each session with a single network setting. Each session is annotated with a satisfaction rating based on a 5-point scale. All the QoS tuples are now classified into 5 different classes based on the rating where the QoS tuples from the same class denotes almost similar quality satisfaction index. Now, the final VoIP experiments are modeled by randomly alternating among the classes, followed by picking one QoS tuple from the selected class and then applying the setting to the VoIP network environment through dummynet pipes in every 20 sec interval cycle. Within each 20 sec interval cycle, the applied QoS values are decayed according to exponential distribution.

C. Design

Another problem of VoIP like network applications is the nature of time-varying quality. We follow the natural approach of dividing the call duration into fixed time intervals (15 seconds) and assess quality of each interval independently in our experiments. Our main goal in this paper is not to study the time-varying quality but to effectively analyze the quality modeling and prediction from an user-centric approach of affective response. We perform acoustic feature extraction from the recorded voice signals with the help of openSMILE [9] tool. We capitalized the state-of-art Automatic Speech Recognition (ASR) system from the HTK Toolkit [14] of Cambridge University which uses Hidden Markov Models (HMM) for predicting text output of voice signals. The trained HMM models are accessed from [13] which are trained on Wall Street Journal (WSJ) corpora and the generated tied-state cross-word triphones are utilized for later recognition purposes. We create a 3-gram language model provided in a 1200 million English Gigaword corpus [2] indexed within the Linguistic Data Consortium (LDC) catalogue and also coupled with a 125,000 word CMU pronunciation dictionary [1]. We exploit libSVM [7] tool for SVM classifiers and Weka [10] for kNN classifiers.

D. Participants and Questionnaires

Fifteen participants from different undergraduate programs were selected for the study through a campus-wide ad. The participants were from 2 different undergraduate classes: Introductory Java Programming and Computer Networks. Among the 15 participants, 11 were male and 4 were female. All participants were between the ages of 18-38 years of age (M=25.54, SD=3.38). They had a mean of 3.67 years of using VoIP service and all claimed to have been using at least one VoIP service in the past. Each participant was asked to complete an Entry Questionnaire at the beginning of the study for collecting background and demographic information and was further inquired about their previous experience with VoIP systems if any. Exit Questionnaires were also supervised at the end of each session to elicit subjective experience of the user for the entire audio conversation. We executed quiz-based conversation sessions where the participants will be heavily engaged with interaction. Quizes are high-voltage cognitive task with formidable degree of reciprocity and zest. For achieving common ground, we designed 2 quiz sessions based on Java programming and Computer Networks where the questions were carefully selected to elicit subjective answers. Moreover, to avoid over-burdening the brain of the participants’, we frequently pose some comments on general topics or short riddles, for providing cognitive relief from course-related information exchange.

E. Results

In this section we study the experimental findings of our study based on 15 subjects in our laboratory environment. For all the type of information sources, we randomly divided the data samples into 75% for training and 25% for testing in general cases. Table: I depicts the classification accuracy between different feature sets for several classification methods such as variants of SVM. In all the cases, the PCA feature set derives the highest performance in terms of classification accuracy. The f10 and f15 feature sets almost generate comparable accuracy percentage which is consistently higher than the base feature set. This proves that the extra computational overhead incurred due to the extraction of f10, f15 and PCA feature sets are contributory. Comparing across different classification methods, we can broadly observe that SVM with 5-fold cross-validation gives the best prediction accuracy among all the other techniques. The boosting performed by hierarchical technique with 5/10 weak classifiers (5WC/10WC) do not provide

<table>
<thead>
<tr>
<th>Classifier Model</th>
<th>Base</th>
<th>f10</th>
<th>f15</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM-Plain</td>
<td>57.3</td>
<td>61.1</td>
<td>62.5</td>
<td>65.7</td>
</tr>
<tr>
<td>SVM-3WC</td>
<td>58.7</td>
<td>62.3</td>
<td>62.7</td>
<td>64.8</td>
</tr>
<tr>
<td>SVM-7WC</td>
<td>58.9</td>
<td>61.5</td>
<td>62.4</td>
<td>65.2</td>
</tr>
<tr>
<td>SVM-5CV</td>
<td>60.6</td>
<td>63.7</td>
<td>64.1</td>
<td>67.9</td>
</tr>
</tbody>
</table>

The next step involves conducting multiple test sessions with different QoS tuples, but each session with a single network setting. Each session is annotated with a satisfaction rating based on a 5-point scale. All the QoS tuples are now classified into 5 different classes based on the rating where the QoS tuples from the same class denotes almost similar quality satisfaction index. Now, the final VoIP experiments are modeled by randomly alternating among the classes, followed by picking one QoS tuple from the selected class and then applying the setting to the VoIP network environment through dummynet pipes in every 20 sec interval cycle. Within each 20 sec interval cycle, the applied QoS values are decayed according to exponential distribution.

C. Design

Another problem of VoIP like network applications is the nature of time-varying quality. We follow the natural approach of dividing the call duration into fixed time intervals (15 seconds) and assess quality of each interval independently in our experiments. Our main goal in this paper is not to study the time-varying quality but to effectively analyze the quality modeling and prediction from an user-centric approach of affective response. We perform acoustic feature extraction from the recorded voice signals with the help of openSMILE [9] tool. We capitalized the state-of-art Automatic Speech Recognition (ASR) system from the HTK Toolkit [14] of Cambridge University which uses Hidden Markov Models (HMM) for predicting text output of voice signals. The trained HMM models are accessed from [13] which are trained on Wall Street Journal (WSJ) corpora and the generated tied-state cross-word triphones are utilized for later recognition purposes. We create a 3-gram language model provided in a 1200 million English Gigaword corpus [2] indexed within the Linguistic Data Consortium (LDC) catalogue and also coupled with a 125,000 word CMU pronunciation dictionary [1]. We exploit libSVM [7] tool for SVM classifiers and Weka [10] for kNN classifiers.
reasonable performance gain compared to the plain technique of training/testing only. The best model is found to be SVM 5-fold cross-validation with PCA-based feature selection which gives 67.9% classification accuracy. Next, we investigate the performance effect by the combination of different information sources (i.e., acoustic, linguistic) with respect to overall classification accuracy. The result is shown in Figure 2 where the combination of sources are: acoustic only, linguistic only, and acoustic+linguistic (same as in Table: I). Some of the overall conclusions are: (1) combining other sources with acoustic information consistently improves performance with the best result reported from the combination of acoustic+linguistic, (2) considering single source, acoustic performs much better than linguistic module.

IV. DISCUSSION & CONCLUSIONS

Implicit or non-intrusive quality evaluation models are appealing due to its lesser resource overhead and completely automated methodology, but on the other hand is very challenging for its subjective non-deterministic behavior. We tried to explore in this direction by evaluating an affective feedback framework for capturing subjective perception of the user and its influence on VoIP QoE evaluation in a controlled experimental framework with human participants. We analyze an affective feedback based framework for the purpose of capturing subjective perception by combining diverse information sources such as acoustic, lexical and discourse features. We evaluated the framework by well known classification techniques such as SVM and kNN which were used to train models capable of discriminating the QoE categories successfully. We derive results showing reasonable classification accuracy and the maximum is achieved by combining all the information sources selected through PCA. The accumulated evidence supports our hypothesis of exploiting affective response as a predictor of subjective experience of quality due to its high correlation with cognitive perception. However, we refrain from claiming that our method can capture the entire spectrum of subjective quality perception, but we provide contributory evidence for affective information to be considered as one relevant indicator. We are aware that our present study has several limitations such as in some cases the content semantics might have influenced affective reactions which were unrelated to quality evaluation. However, we feel that this issue can be dealt in the future by considering multi-modal knowledge sources.

Some of the open issues that need to be further explored which are possible avenues for future work are as follows: (1) influence of other affective cues such as laughter, sigh, etc. on the subjective quality of experience, (2) integrating other discourse related attributes such as rephrase, reject, start-over in the present framework for further evaluation and the problem of automatically detecting these indicators from voice signals, (3) effective classifier combination/fusion techniques with more accurate results and additional pattern recognition mechanisms for optimising the extraction of affective information from voice signals, (4) comparing other implicit feedback approaches in the context of QoE evaluation considering different levels of interactivity for real-time collaborating applications such as VoIP.

REFERENCES